American
Psychological
Association

Volume 110
Number 1

January 2018

Published eight times

ISSN 0022-0663

Journal of Educational Psychology

www.apa.org/pubs/journals/edu

Editor

Associate Editors

Consulting Editors

The main purpose of the *Journal of Educational Psychology*® is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Single Issues, Back Issues, and Back Volumes: For information regarding single issues, back issues, or back volumes, write to Order Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242; call 202-336-5600 or 800-374-2721; or visit www.apa.org/pubs/journals/subscriptions.aspx

Manuscripts: Submit manuscripts electronically through the Manuscript Submissions Portal found at www.apa.org/pubs/journals/edu according to the Instructions to Authors found elsewhere in this issue (see table of contents). Correspondence regarding manuscripts should be sent to the Editor, Steve Graham, at steve.graham@asu.edu.

Disclaimer: APA and the Editors of *Journal of Educational Psychology* assume no responsibility for statements and opinions advanced by the authors of its articles.

Electronic Access: APA members who subscribe to this journal have automatic access to all issues of the journal in the PsycARTICLES® full-text database. See http://my.apa.org/access.html.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

APA Journal Staff: Rosemarie Sokol-Chang, PhD, *Publisher, APA Journals*; Mare Meadows, *Managing Director*; Richard Pepple, *Journal Production Editor*; Cheryl Johnson, *Peer Review Coordinator*; Jodi Ashcraft, *Director, Advertising Sales and Exhibits.*

Journal of Educational Psychology® (ISSN 0022-0663) is published eight times (January, February, April, May, July, August, October, November) in one volume per year by the American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Subscriptions are available on a calendar year basis only (January through December). The 2018 rates follow: *Nonmember Individual:* $263 Domestic, $305 Foreign, $327 Air Mail. *Institutional:* $1,020 Domestic, $1,097 Foreign, $1,121 Air Mail. *APA Member:* $127. *APA Student Affiliate:* $79. Write to Subscriptions Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Printed in the U.S.A. Periodicals postage paid at Washington, DC, and at additional mailing offices. POSTMASTER: Send address changes to *Journal of Educational Psychology,* American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

Effective with the 1986 volume, this journal is printed on acid-free paper.

Journal of

Educational Psychology®

www.apa.org/pubs/journals/edu

January 2018

Volume 110
Number 1

Social Relations

Assessment

The Three R's

Early Learners

(*Contents continue*)

Other

Journal of Educational Psychology
2018, Vol. 110, No. 1, 1–16

0022-0663/18/$12.00 http://dx.doi.org/10.1037/edu0000191

A Cluster Randomized Trial of the Social Skills Improvement System-Classwide Intervention Program (SSIS-CIP) in First Grade

James Clyde DiPerna, Puiwa Lei, Weiyi Cheng, Susan Crandall Hart, and Jillian Bellinger
The Pennsylvania State University

The purpose of this study was to evaluate the efficacy of a universal social skills program, the Social Skills Improvement System Classwide Intervention Program (SSIS-CIP; Elliott & Gresham, 2007), for students in first grade. Classrooms from 6 elementary schools were randomly assigned to treatment or business-as-usual control conditions. Teachers assigned to the treatment condition implemented the SSIS-CIP over a 12-week period. Students' social skills, problem behaviors, and approaches to learning were assessed via teacher ratings and direct observations of classroom behavior. In addition, their early literacy and numeracy skills were measured via computer-adaptive standardized tests. SSIS-CIP participation yielded small positive effects in students' social skills (particularly empathy and social engagement) and approaches to learning (academic motivation and engagement). Students' problem behaviors and academic skills, however, were unaffected by SSIS-CIP exposure.

Educational Impact and Implications Statement

The purpose of this study was to evaluate student outcomes associated with a classroom social skills program, the Social Skills Improvement System Classwide Intervention Program (SSIS-CIP; Elliott & Gresham, 2007). Participation in the SSIS-CIP yielded small positive effects in first grade students' empathy, social engagement, academic motivation, and academic engagement. Students' problem behaviors and academic skills, however, were unaffected by SSIS-CIP exposure. Although some outcomes were similar to an earlier study of the SSIS-CIP in second grade classrooms, the first grade findings were consistently smaller in magnitude. If these findings are replicated in future studies, educators and administrators contemplating adoption of the SSIS-CIP should consider prioritizing second grade for implementation of the program within the primary grades.

Keywords: SSIS-CIP, social skills, approaches to learning, social emotional learning, cluster randomized trial

Supplemental materials: http://dx.doi.org/10.1037/edu0000191.supp

The development of social-emotional competence is critical for young children's later success and well-being. Researchers have identified a number of key social-emotional skills within the school setting such as maintaining positive relationships with peers and adults, social problem solving, effectively communicating emotions, listening, and being attentive (Ashdown & Bernard, 2012; DiPerna, Volpe, & Elliott, 2005; Shonkoff & Philips, 2000). The National Academy of Sciences reported that 60% of children enter school with the cognitive skills needed to be successful, but only 40% have the social-emotional skills to succeed (Ashdown & Bernard, 2012). Evidence-based universal interventions focused on social, emotional, and academic competence represent a promising approach to promoting positive youth development (Bradshaw, Zmuda, Kellam, & Ialongo, 2009). Fostering students' social-emotional learning (SEL) using such approaches has been shown to improve both social and academic outcomes (Ashdown & Bernard, 2012).

There is a substantive body of research linking the development of social—emotional competence with school success (Blair & Raver, 2015; Denham et al., 2003; Miles & Stipek, 2006; Wentzel & Asher, 1995). Within the classroom, positive social-emotional skills enable children to develop constructive relationships with teachers and peers, accompanied by foundational learning-related attitudes and behaviors that allow them to become engaged in the many new tasks put before them (Denham, Way, Kalb, Warren-Khot, & Bassett, 2013). Children who demonstrate behavior in a manner consistent with classroom expectations, engage with in-

This article was published Online First March 23, 2017.

James Clyde DiPerna, Puiwa Lei, Weiyi Cheng, Susan Crandall Hart, and Jillian Bellinger, Department of Educational Psychology, Counseling, and Special Education, The Pennsylvania State University.

Jillian Bellinger is now a licensed psychologist in Austin, Texas.

James Clyde DiPerna is the lead author of the Academic Competence Evaluation Scales and receives a royalty from the publisher (Pearson).

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A090438 to The Pennsylvania State University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Correspondence concerning this article should be addressed to James Clyde DiPerna, Department of Educational Psychology, 105 CEDAR Building, The Pennsylvania State University, University Park, PA 16802. E-mail: jdiperna@psu.edu

struction, and persist with learning tasks exhibit higher levels of achievement in school (McClelland, Acock, & Morrison, 2006). Conversely, attention problems undermine effective learning and contribute to off-task behavior and reduced achievement (Hughes & Kwok, 2006). As such, social and behavioral competencies may be as important for children's later success as their early cognitive skills (Bub, 2009; Heckman, 2006).

Universal SEL and Students' Social, Emotional, and Academic Development

Universal intervention programs implemented within a classroom can complement academic instruction to promote children's social, emotional, cognitive, and academic skill development (Blair & Raver, 2015; Greenberg, Domitrovich, & Bumbarger, 2001; Linares et al., 2005). For example, Bierman et al. (2008) implemented a social-emotional and literacy intervention in Head Start classrooms that yielded moderate effects on students' literacy skills and social cognitions. Similarly, the Positive Action program (Flay, Allred, & Ordway, 2001) has demonstrated positive effects on students' behavior, school involvement, and achievement (Flay & Allred, 2003). With regard to long-term positive outcomes, implementation of the Good Behavior Game (Barrish, Saunders, & Wolf, 1969) and an enhanced academic curriculum in first grade resulted in higher scores on standardized achievement tests in twelfth grade, higher rates of high school graduation, and higher rates of college attendance (Bradshaw et al., 2009).

Beyond these studies of individual programs, there have been several meta-analyses of outcomes associated with SEL programs during the past decade. Nelson, Westhues, and MacLeod (2003) completed a meta-analysis of 34 universal SEL programs and reported small-to-moderate positive effects on K-8 students' cognitive development, social-emotional behavior, and parent-family wellness. More recent meta-analyses have provided support for the effectiveness of universal SEL programs for students as well. For example, January, Casey, and Paulson (2011) reported a small mean overall effect of classwide social skills interventions with a wide range of effects reported across studies. In a broader meta-analysis of 213 school-based, universal SEL programs, Durlak, Weissberg, Dymnicki, Taylor, and Schellinger (2011) reported small-to-moderate effects on students' positive social behavior in daily situations, as rated by students, teachers, parents, or independent observers. The study also reported larger mean effects for students' social-emotional skill performance, which included skills assessed via hypothetical scenarios, test situations, or structured tasks but did not include teacher ratings of students' behaviors in daily classroom settings. Finally, Durlak et al. (2011) also reported positive effects relative to students' academic performance, emotional distress, and conduct problems.

SEL Programs and the Primary Grades

As children move through the primary grades, they are establishing key social-emotional (e.g., self-regulation, peer relationships) and academic (e.g., letter naming, vocabulary skills) competencies necessary for later school success (Blair & Raver, 2015; Early, Pianta, & Cox, 1999). As such, universal programs incorporating evidence-based instruction and learning activities characterized by systemic, direct, intentional instruction are critical for supporting children's social, emotional, and educational outcomes during this critical developmental period (Bub, 2009; Durlak, Weissberg, & Pachan, 2010; Nelson et al., 2003). The Social Skills Improvement System-Classwide Intervention Program (SSIS-CIP; Elliott & Gresham, 2007) is a universal program developed for use within the general education classroom, and the program utilizes instructional strategies (e.g., reinforcement, modeling, role-playing, and problem-solving) grounded in several established theories of student learning (e.g., operant, social learning). The curriculum includes 10 instructional units targeting social-emotional skills. Each unit targets one specific skill (e.g., listening to others, asking for help, and getting along with others) and includes three brief lessons. Across the 10 skill units, five focus on cooperation skills, two feature self-control skills, and the remaining three units focus on assertion, responsibility, and empathy. (Additional details regarding the SSIS-CIP are provided in the Method section.)

Because the SSIS-CIP targets self-regulatory behaviors that have been shown to complement and enhance learning in classroom settings (e.g., Blair & Raver, 2015), its underlying theory of change postulates proximal, medial, and distal student outcomes resulting from exposure to the program. Specifically, proximal outcomes include improvement in the social and emotional skills explicitly taught within the SSIS-CIP curriculum. Medial outcomes consist of improvements in students' approaches to learning (academic motivation and engagement) and reductions in problematic classroom behaviors (e.g., acting out, inattention). Distal outcomes are positive changes in academic skills. The SSIS-CIP theory of change is not only grounded in theoretical models linking student behavior, approaches to learning, and academic achievement (e.g., DiPerna et al., 2005) but also consistent with the SEL outcomes framework specified by the Collaborative for Academic, Social, and Emotional Learning (2016).

Two other popular universal SEL programs that, like SSIS-CIP, focus on the promotion of social skills and academic readiness behaviors in young children are The Incredible Years Classroom Dinosaur Social Skills and Problem-Solving Curriculum (IYCD; Webster-Stratton & Reid, 2004) and Second Step (Committee for Children, 1992). An adaptation of a popular small-group clinic-based intervention, IYCD is designed for students ages 3–8; Second Step and SSIS-CIP are available in versions from preschool to early adolescence. Though all three curricula have secondary aims of reducing problem behaviors, they approach SEL through different curricular foci. IYCDs approach is broad, including lessons in anger management, problem solving, emotion language, and friendship skills, while Second Step emphasizes the importance of self-regulation, empathy, and problem-solving. In contrast, SSIS-CIP's curriculum focuses on specific discrete social skills identified by a nationally representative sample of teachers as critical for classroom success (Elliott & Gresham, 2007). All three programs are designed to be taught by general education teachers using a variety of instructional strategies including direct instruction, modeling, and role play. IYCD requires the most instructional time to complete (60 lessons of approximately 35–40 min each), while the early elementary versions of Second Step (22 lessons, 25–40 min each) and SSIS-CIP (30 lessons, 20–25 min each) are less time-intensive.

Several randomized trials have been published to date examining the efficacy of these three early elementary programs. With regard to IYCD, Webster-Stratton, Jamila Reid, and Stoolmiller (2008) reported significant improvements in students' (preschool – Grade 1) emotional self-regulation skills, social competence, and conduct problems based on ratings by independent observers. Results demonstrated a pattern of differential effectiveness; specifically, students from classrooms with the most initial risk (i.e., lower school readiness skills and higher conduct problems at baseline) benefited most from the intervention (Webster-Stratton et al., 2008). Furthermore, IYCD was associated with statistically significant increases in preschool students' appropriate behavior, interest, and enthusiasm based on postobservation ratings of whole-class behavior (Baker-Henningham, Walker, Powell, & Gardner, 2009). It is important to note that in both of these trials IYCD was implemented in conjunction with the Incredible Years teacher training program, so the observed outcomes reflect the cumulative effects of both approaches.

Results from two randomized controlled trials indicate mixed support for Second Step's efficacy in early elementary classrooms, with outcomes varying across measurement methods (i.e., behavior ratings vs. direct observation). Grossman et al. (1997) found no significant differences for parent and teacher behavior ratings for second- and third-grade students assigned to intervention and control conditions, but reported significant decreases in observed physical aggression and increases in neutral/prosocial behavior for the Second Step group. In a large trial involving students in kindergarten through second grade (Low, Cook, Smolkowski, & Buntain-Ricklefs, 2015), Second Step yielded marginally significant small main effects in reducing behavior problems, increasing social-emotional skills, and improving skills for learning. However, moderation analyses indicated that effects varied by pretest scores. Specifically, students exhibiting more severe problem behaviors or lower levels of social-emotional skills at pretest demonstrated larger positive gains from Second Step participation than their peers with moderate or higher levels of initial skills (Low et al., 2015).

Efficacy of the SSIS-CIP has been tested in one randomized trial to date (DiPerna, Lei, Bellinger, & Cheng, 2015, 2016). This trial included 432 second-grade students, 38 classrooms, and multiple outcome measures (including teacher report and direct observation) consistent with the SSIS-CIP theory of change. Results indicated that second-grade students demonstrated greater gains in teacher ratings of overall social skills, communication, cooperation, responsibility, and empathy upon completion of the SSIS-CIP (Early Elementary Level) than students who were not exposed to the program (DiPerna et al., 2015). Although students in SSIS-CIP classrooms demonstrated fewer withdrawn behaviors, other internalizing and externalizing behaviors did not demonstrate statistically significant differences. In addition, students exposed to the SSIS-CIP demonstrated improvement in their academic motivation and engagement relative to their peers in nonimplementing (control) classrooms (DiPerna et al., 2016). Improvement in academic skills relative to peers, however, was only observed in mathematics for students receiving supplemental (Title 1) services. Similar to the Webster-Stratton et al. (2008) study of IYCD and Low et al. (2015) study of Second Step, the impact of the SSIS-CIP was moderated by pretest skill level with students demonstrating lower skills at pretest (social skills, academic motivation, and engagement) benefitting most from the program.

Purpose, Rationale, and Hypotheses

Although the SSIS-CIP Early Elementary Level curriculum was developed for Grades 1–3 and is being implemented in elementary schools across the United States, there has only been one study of its efficacy to date. In addition, this study focused solely on students in second grade, and outcomes may not generalize to younger students as a result of age-related differences in social and cognitive functioning. Given the importance (as well as paucity) of systematic replication in educational and psychological science (e.g., Makel & Plucker, 2014; Makel, Plucker, & Hegarty, 2012), the purpose of this study was to examine the efficacy of the SSIS-CIP in first grade classrooms. Specifically, we tested hypotheses informed by the SSIS-CIP theory of change, initial efficacy trial in second grade, and results of studies of other universal programs in the primary grades (Webster-Stratton et al., 2008; Low et al., 2015). The first hypothesis was that children in classrooms implementing the SSIS-CIP demonstrate improved social skills compared to children in nonimplementing (business-as-usual) control classrooms. Second, children in the SSIS-CIP condition were expected to demonstrate fewer problem behaviors than their peers in control classrooms. The third hypothesis was that children exposed to the SSIS-CIP demonstrate improved approaches to learning, and the fourth hypothesis was that SSIS-CIP students demonstrate improved academic skills relative to their peers. Finally, given the findings of the Grade 2 trial as well as Low et al.'s (2015) Second Step efficacy trial, we examined if SSIS-CIP effects with first grade students are moderated by their initial skill level in the target outcome variable (individually or at the class level.).

Method

Participants

Participating classrooms were drawn from six elementary schools in the Mid-Atlantic region of the United States. All first grade teachers in these six schools ($N = 61$) were invited to participate in the study; however, two teachers were unable to participate because of extended absences resulting from medical or family leave. As such, the total number of participating classrooms at the beginning of data collection was 59 (see Figure 1). Approximately half of these classrooms were from four schools in a small urban school district, and the remaining classrooms were from two elementary schools in a small rural district. Most of the classroom teachers reported extensive classroom experience ($M = 21.81$ years, $SD = 9.50$), and all were White and female.

All students from each participating classroom were invited to participate in the data collection associated with the efficacy trial, and approximately 56% ($N = 766$) received parental consent (see Figure 1). Thirty of the participating students (3.9%) moved before posttest data were collected, and 17 of these students were in the SSIS-CIP condition. An additional 40 students (5.2% of the initial participating student sample) were excluded from the final analyses because of missing data, and 32 of these students were from two classrooms where participating teachers were unable to com-

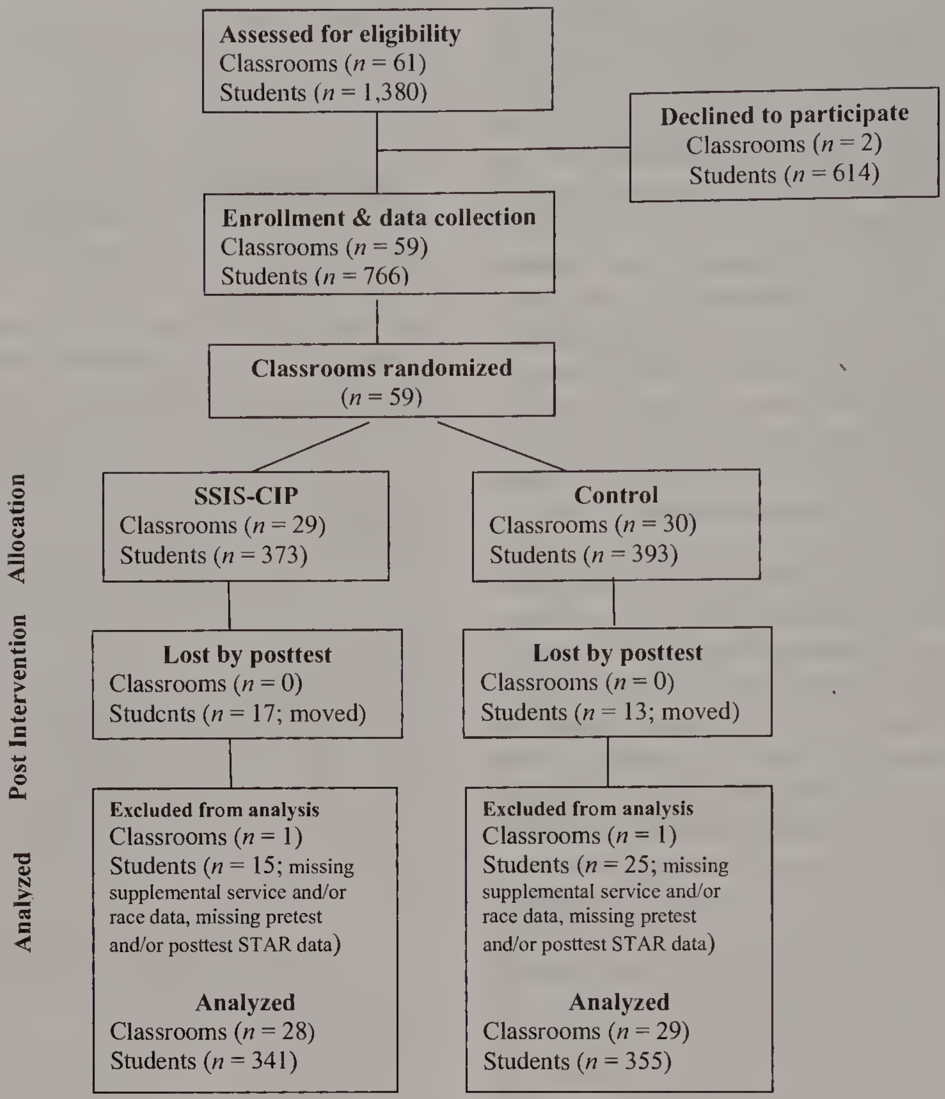


Figure 1. Participant flowchart.

plete the final round of data measures at the end of the academic year.

Demographic characteristics of the analyzed participant sample (see Table 1) were consistent with the first grade student population across the participating elementary schools. The mean age of the overall student sample was 6.29 years, and about half of students were male (53.3%). The racial composition of the student sample included White (70.1%) Black/African American (24%), Hawaiian-Pacific (0.4%), and Asian (4.9%), and approximately 9.4% of students also identified as Hispanic or Latino. A majority of students (94.1%) spoke English as their primary language. Students receiving special education services (via an Individualized Education Program) comprised 6.2% of the total sample, while 27.7% of students received academic support through supplemental services (i.e., Title 1). A small number of students (1.9%) were repeating first grade while enrolled in the study. None of the slight observed differences across demographic characteristics of the SSIS-CIP and control subsamples were statistically significant. All student and teacher participants were treated in accord with the ethical principles of the American Psychological Association.

Table 1
Student Demographic Characteristics by Treatment Condition

Variable	SSIS-CIP (*N* = 341)	Control (*N* = 355)
Age (in years)	6.29 (.42)	6.30 (.43)
Male	51.61	54.93
White	72.43	67.89
Black/African American	21.54	26.07
Hawaiian-Pacific	.90	0
Asian	4.42	5.28
Hispanic or Latino	9.21	9.51
Other race	5.36	3.09
English as primary language	94.72	93.52
Special education	4.40	7.89
Supplemental services	26.98	28.45
Repeating first grade	1.50	2.30

Note. SSIS-CIP = Social Skills Improvement System Classwide Intervention Program. Mean (*SD*) are reported for Age and percentage is reported for all other variables. There were no statistically significant differences ($p < .05$) on any of the demographic variables.

Measures

The measures used to assess the primary student outcomes of interest (social skills, problem behavior, approaches to learning, and academic skills) were identical to those used in the Grade 2 efficacy trial (DiPerna et al., 2015, 2016). Specifically, participants' social skills and problem behaviors were measured via teacher ratings on the Social Skills Improvement Rating Scales-Teacher Form (SSIS-RST; Gresham & Elliott, 2008) and direct observations using the Cooperative Learning Observation Code for Kids (CLOCK; Volpe & DiPerna, 2010). Participants' approaches to learning (academic motivation and engagement) were measured via teacher ratings on the Academic Competence Evaluation Scales (ACES; DiPerna & Elliott, 2000) and direct observation on the CLOCK. Students' academic skills were assessed via the STAR Reading and Math computerized adaptive tests (Renaissance Learning, 2009, 2010). Finally, the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008) was used to assess the instructional environment in each participating classroom before SSIS-CIP implementation.

Social skills and problem behavior. The SSIS-RST (Gresham & Elliott, 2008) was used to obtain teachers' perspectives of their students' social skills and problem behaviors in the classroom setting. The SSIS-RST Social Skills scale includes 46 items, seven subscales (Communication, Cooperation, Assertion, Responsibility, Empathy, Engagement, and Self-Control), and a total composite. The Problem Behaviors Scale includes 24 items across five subscales (Externalizing, Bullying, Hyperactive-Inattentive, Internalizing, and Autistic Behavior); however, the Autistic Behavior subscale was not analyzed in the current study. Teachers rate each item on the Social Skills and Problem Behaviors scales using a 4-point format ranging from *never* to *almost always*. Psychometric evidence for SSIS-RST scores is consistent with its intended purpose (Gresham & Elliott, 2008), and reliability indices (α = .88–.98) in the current sample are strong (see Table 2).

The CLOCK (Volpe & DiPerna, 2010) is a structured observation protocol that was used to facilitate independent direct observations of student social and problem behavior in the classroom. Specifically, the CLOCK category of Positive Social reflects any social behavior that is permitted during the observation interval, and Interference measures student problem behaviors that distract others or disrupt the functioning of the classroom. Each of these behaviors is observed using a partial interval format with each interval lasting 15 s. Six participants (three boys and three girls) were randomly identified within each classroom, and each was observed on three separate occasions during each data collection window. All direct observations were completed during mathematics to standardize the instructional context. (Both participating districts used Everyday Math curriculum, which features collaborative learning.)

Observers (*N* = 39) had at least a bachelor's degree in psychology, education, or a related discipline. In addition, they completed formal training regarding the CLOCK (approximately 12 hr of didactic instruction, practice observations, and individualized feedback) and had to meet a mastery criterion (80% accuracy when observing a video of students in an elementary classroom) before they could conduct observations for the project. Observations were distributed approximately evenly across observers, and each observation lasted for 12 min. One-third of the CLOCK observations were completed by pairs of observers, and agreement was high (κ = .88–.94) across all target behavior domains and paired observations (see Table 2).

Approaches to learning. The ACES (DiPerna & Elliott, 2000) was used to measure teacher perspectives regarding their students' approaches to learning (academic motivation and en-

Table 2
Reliability Indices and Intraclass Correlations for Social Skills, Problem Behaviors, Approaches to Learning, and Academic Skills

	Reliability index		ICC (school)		ICC (class)	
Variable	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest
Social skills						
Social skills composite	.98	.98	.07	.07	.18	.16
Communication	.91	.91	.04	.04	.20	.25
Cooperation	.94	.94	.08	.07	.10	.06
Assertion	.88	.88	.09	.14	.23	.24
Responsibility	.92	.91	.06	.05	.13	.08
Empathy	.93	.95	.01	.04	.24	.21
Social engagement	.93	.94	.05	.04	.19	.20
Self-control	.93	.94	.07	.07	.15	.13
Positive social[a]	.88	.91	.01	.005	.15	.14
Problem behaviors						
Externalizing	.94	.94	.03	.01	.19	.16
Bullying	.90	.92	.04	.01	.20	.17
Hyperactive-inattentive	.90	.91	.06	.06	.15	.13
Internalizing	.89	.88	.04	.02	.32	.37
Interference[a]	.89	.93	.03	.01	.20	.27
Approaches to learning						
Academic motivation	.98	.98	.04	.07	.12	.09
Academic engagement	.96	.95	.04	.07	.11	.13
Engaged time[a]	.92	.94	.04	.08	.28	.32
Academic skills						
Math scaled score	—	—	.14	.11	.09	.09
Reading scaled score	—	—	.06	.09	.07	.07

Note. ICC = Intraclass correlation. Reliability indices are Cronbach's αunless noted otherwise.
[a] Direct observation data (κ agreement index reported for reliability).

gagement) in the classroom. The ACES Academic Motivation subscale includes 11 items that measure a student's approach, persistence, and level of interest regarding academic learning. The Academic Engagement subscale includes eight items that reflect attention and active participation in classroom activities. Items are rated using a 5-point format ranging from *never* to *almost always*. Psychometric evidence for ACES scores is consistent with its intended purpose (DiPerna & Elliott, 2000), and reliability estimates (α = .95–.98) from the current sample are strong (see Table 2).

Direct observation of students' engagement during instruction was also completed as part of the aforementioned CLOCK observations. The CLOCK category of Engaged Time includes both active (e.g., raising hand, asking teacher a relevant question) and passive engagement (e.g., listening to a teacher talk, looking at the whiteboard or a worksheet) in classroom instruction. Engaged time is observed using a partial interval format with each interval lasting 15 s.

Academic skills. The STAR Math (Renaissance Learning, 2009) and Reading (Renaissance Learning, 2010) computerized adaptive tests were used to directly measure students' academic skills. STAR Math is composed of a series of multiple choice mathematical problems that assess proficiency with numeration and computation objectives. STAR Reading features vocabulary-in-context items that require students to utilize background information, apply vocabulary, and use active strategies to construct meaning. Each STAR assessment requires approximately 10 min to complete, and both were administered before and after intervention implementation. STAR scores demonstrate high reliability and strong relations with other standardized achievement test scores (e.g., CA Achievement Test, Stanford Achievement Test) as well as teacher ratings of students' academic proficiency (Renaissance Learning, 2009, 2010).

Classroom instructional environment. Each participating classroom was observed once during the first data collection window to determine if there were significant differences in instructional environments across the participating classrooms, and if so, control for them in the tests of the hypotheses. The CLASS K-3 (Pianta et al., 2008) is a structured observation system that yields scores in three domains: Emotional Support, Classroom Organization, and Instructional Support. These broad domains are further differentiated across 10 dimensions (Positive Climate, Negative Climate, Teacher Sensitivity, Regard for Student Perspectives, Behavior Management, Productivity, Instructional Learning Formats, Concept Development, Quality of Feedback, and Language Modeling). Each dimension is rated on a 7-point scale ranging from *low* to *high*. Ratings are assigned after an observer completes an observation "cycle" (20 min of observation followed by 10 min of assigning ratings to dimensions/domains). Psychometric evidence for the CLASS is sound (Hamre, Mashburn, Pianta, & LoCasale-Crouch, 2008) and provides support for its intended purpose.

Observers were formally trained by a CLASS-certified instructor and achieved the CLASS-mastery criterion (>80% accuracy) before completing observations. Consistent with the authors' recommendations, two observation cycles were completed in each classroom to yield representative dimension and domain scores. Domain scores demonstrated acceptable levels of internal consistency (.81–.93). In addition, paired observations were completed for approximately 33% of the classrooms (N = 21), and intraclass correlations between these paired observations were moderate to high (.65–.76) for the CLASS domain scores.

Procedure

Recruitment. Data were collected as part of a multiyear project examining efficacy of the SSIS-CIP. First grade teachers within each elementary school were invited to participate in the project. Upon receiving teachers' consent, letters were distributed to parents requesting consent for their child's participation in the data collection process. A reminder letter was sent approximately four school days after the initial invitation letter.

Data collection. Both the business-as-usual control and treatment classrooms followed the same data collection schedule. Child-level data were collected during 4-week periods before (November–December) and after (March–April) SSIS-CIP implementation (average pre- and posttest interval = 4.8 months). Specifically, teachers completed the SSIS-RST and ACES subscales for all participating children. All participating teachers were compensated for time spent completing these scales. In addition, trained examiners administered the STAR assessments to all students with parental consent. (Verbal assent also was obtained from students before any testing commenced.) Research staff also completed CLOCK observations for the randomly selected subsample (3 boys and 3 girls) from each classroom. As noted previously, each student was observed during mathematics instruction on three separate days during the pre- and postdata collection periods. (If a student was absent on a scheduled observation day, the observer rescheduled the observation for a mathematics period later in the week.)

SSIS-CIP. The SSIS-CIP includes 10 instructional units focused on key classroom social skills identified by teachers as important for classroom success. Specifically, Units 1–3 target receptive skills (i.e., listening to others, following the steps, following the rules), Unit 4 focuses on selective input (i.e., paying attention to your work), Unit 5 focuses on productive skills (i.e., asking a question), and Units 6–10 target interactive skills (i.e., communicating, cooperating, reading or managing emotions, and showing an understanding of rules). Each unit includes three scripted lessons, brief video vignettes (30–90 s), and practice exercises (student booklets). Each lesson requires approximately 20–25 min to complete and relies on six instructional strategies (describe, model, role-play, do, practice, monitor progress, and generalize) to help children learn the target skill for that unit. Additional information regarding the SSIS-CIP is available in the Instructor's Handbook (Elliott & Gresham, 2007).

Intervention teachers (N = 29) completed a 1-day workshop with the lead author before implementation. The first half of the workshop provided a detailed overview of the SSIS-CIP lesson plans, student booklets, and video vignettes. During the second half, participants practiced teaching each lesson from the first SSIS-CIP unit in small groups. As teachers practiced, the workshop facilitator provided structured feedback regarding fidelity of their role-play lessons. In addition, teachers had the opportunity to ask questions regarding curricular implementation. After the formal training, implementing teachers were expected to teach one SSIS-CIP unit (three lessons) per week.

Two complimentary methods were used to evaluate and ensure fidelity of implementation of the SSIS-CIP lessons. First, research staff completed direct observations for approximately 20% of the SSIS-CIP lessons taught in each classroom. Observations were completed approximately every other week to ensure lessons were sampled throughout the implementation period. For each fidelity observation, staff observed an entire lesson and then completed a structured report form that included 20 specific instructional actions/activities. Observers recorded if each activity was completed (or not) during the observed lesson and provided a summative judgment regarding the overall implementation of the five core lesson components (introduce, define, discuss, identify steps and practice, and model/role-play) using a 4-point scale ranging from *not implemented* (1) to *full implementation* (4). In addition, implementing teachers completed weekly standardized checklists indicating their level of implementation (using the same 4-point scale as the observers) for the five core components of each lesson.

During the implementation period, fidelity (both self-report and independent observations) was monitored to ensure that teachers demonstrated at least 90% fidelity in their implementation of the SSIS-CIP lessons, and all SSIS-CIP teachers consistently met this threshold throughout the implementation period. In addition, the research team periodically checked with all teachers (approximately every other week) to see if they had any implementation questions, make sure nothing had arisen that would adversely impact implementation of the SSIS-CIP lessons, and thank them for their efforts. As a result of the scripted lesson format and these monitoring efforts, the SSIS-CIP program was fully implemented across all classrooms based on summative ratings by teachers (M = 3.92, SD = .16) and independent observers (M = 3.97, SD = .08).

Control condition. Teachers in classrooms randomly assigned to the control condition (N = 30) continued with their daily approach to managing and promoting positive classroom behavior throughout the duration of the study. Our primary reason for choosing a business-as-usual control condition is that most schools considering use of the SSIS-CIP likely would be doing so because their teachers do not have a similar program already in place. Thus, understanding magnitude of effect relative to typical classroom practices (i.e., locally developed behavior management plans) would be helpful to stakeholders when making decisions about adopting the SSIS-CIP. Approximately 85% of the control teachers reported having an explicit, planned approach to promoting positive behavior in their classroom. These plans primarily focused on the use of reinforcement (verbal praise or systems where students earned points for positive behavior toward tangible rewards) and/or consequences in the classroom. No teacher in the control condition reported use of a structured curriculum focused on the instruction of social skills.

Design and Data Analyses

This study used a multisite Cluster Randomized Trial (CRT) to test the efficacy of SSIS-CIP on each of the key outcome variables. Classrooms were randomly assigned to experimental conditions (SSIS-CIP and business-as-usual control) within schools. Figure 1 depicts the flow of classroom and student participants throughout the study. There was a low percentage of missing data (0.8–4.4% across variables), and these data were missing completely at random (Little's χ^2 = 496.76, df = 529, p > .05). Given listwise deletion yields unbiased estimates under these conditions (Graham, 2009), cases were deleted listwise for analysis.

Multilevel modeling was used to account for the data structure of students nested within classes (M = 12 students/class, SD = 4). We first estimated unconditional models to report intraclass correlation (ICC) coefficients that indicated the degree to which the assumption of independence was violated because of the clustering of students in classes in schools (Raudenbush, 1997). Table 2 presents ICCs at both school- and class-levels for all outcome measures. Class-level ICCs for posttest outcome measures ranged from small (.06 for Cooperation) to large (.37 for Internalizing). These levels of ICCs suggested that standard errors might be underestimated if the nested data structure was not taken into account. Therefore, we analyzed a two-level model for each outcome to provide proper SE estimates. School-level variances of all posttest outcome measures were mostly small and statistically nonsignificant based on z tests (two-tailed ps > .05). However, school-level ICCs for several posttest measures (e.g., .14 for Assertion, .11 for Mathematics) were considered medium-sized (Raudenbush, Spybrook, Liu, & Congdon, 2005). Therefore, we included school indicators (dummy-coded) to control for school differences in all two-level models.

As recommended by What Works Clearinghouse (U.S. Department of Education, What Works Clearinghouse, 2016), we also tested for baseline equivalence between treatment and control conditions with respect to student demographic variables, classroom quality (i.e., CLASS scores), and all pretest measures. We then controlled for variables that showed nonequivalence when evaluating treatment effects. There were no statistically significant differences in CLASS scores between treatment and control classrooms. Moreover, there were no statistically significant differences between treatment and control conditions (based on two-level models) on pretest measures of problem behaviors, academic motivation, academic engagement, and academic skills.

Children in the control condition, however, had higher pretest scores on social skill measures than children in SSIS-CIP condition (Tables 3 and 4). As such, this baseline nonequivalence between conditions was addressed by including the social skills composite pretest (grand-mean centered) as an additional covariate for all outcome measures, including social skill subscales. Although social skill subscale scores were expected to be correlated with the composite, there was no substantial collinearity when the composite pretest was included as a covariate in the social skill subscale models. This was likely because each subscale (as one of seven) only contributed a small fraction to the composite variance. Because of the importance of statistically adjusting for nonequivalent baseline characteristics in the evaluation of treatment effects (U.S. Department of Education, What Works Clearinghouse, 2016) and the lack of multicollinearity, we decided to include the social skill composite pretest as a covariate even for social skill subscale outcomes.

To address the primary research questions regarding SSIS-CIP outcomes, we included both student- and class-level predictors to adjust for their effects. Student-level predictors included pretest scores of social skill composite (grand-mean centered) and the respective outcome measure (group-mean centered), students' gender (1 = male, 0 = female), race (1 = White, 0 = other), receipt of special education services (1 = yes, 0 = no), and receipt of

Table 3
Student-Level Means (SDs) by Measure, Time, and Treatment Condition

	Pretest		Posttest	
Variable	SSIS-CIP	Control	SSIS-CIP	Control
Social skills				
Social skills composite	1.92 (.51)	2.18 (.50)	2.20 (.51)	2.31 (.49)
Communication	2.00 (.62)	2.31 (.54)	2.29 (.56)	2.46 (.52)
Cooperation	1.90 (.67)	2.12 (.67)	2.13 (.68)	2.22 (.66)
Assertion	1.70 (.58)	1.94 (.60)	2.01 (.61)	2.14 (.57)
Responsibility	2.00 (.63)	2.24 (.61)	2.23 (.61)	2.33 (.60)
Empathy	1.97 (.58)	2.27 (.57)	2.31 (.60)	2.38 (.60)
Social engagement	2.00 (.59)	2.25 (.57)	2.30 (.56)	2.41 (.57)
Self-control	1.88 (.59)	2.11 (.62)	2.13 (.59)	2.21 (.63)
Positive social[a]	.36 (.69)	.39 (.67)	.31 (.58)	.28 (.52)
Problem behaviors				
Externalizing	.61 (.57)	.50 (.53)	.57 (.57)	.47 (.54)
Bullying	.34 (.50)	.27 (.46)	.34 (.52)	.24 (.45)
Hyperactive-inattentive	.89 (.67)	.78 (.65)	.80 (.67)	.69 (.64)
Internalizing	.55 (.55)	.51 (.51)	.51 (.53)	.46 (.49)
Interference[a]	.17 (.27)	.27 (.49)	.18 (.35)	.33 (.68)
Approaches to learning				
Academic motivation	3.28 (1.06)	3.42 (1.01)	3.60 (1.00)	3.62 (1.05)
Academic engagement	3.51 (1.02)	3.66 (.96)	3.91 (.88)	3.91 (.91)
Engaged time[a]	74.73 (13.92)	79.50 (13.75)	78.06 (14.49)	79.61 (13.99)
Academic Skills				
Math scaled score	257.00 (153.04)	250.95 (151.80)	356.05 (128.45)	351.32 (132.88)
Reading scaled score	88.35 (90.76)	81.41 (76.59)	152.10 (106.55)	140.10 (96.94)

Note. SSIS-CIP = Social Skills Improvement System Classwide Intervention Program. SSIS-CIP N = 341; control N = 355 (unless noted).
[a] Direct observation data (SSIS-CIP N = 157; control N = 161).

supplemental (Title 1) services (1 = yes, 0 = no). The dummy variable predictors were grand-mean centered. Class-level predictors included grand-mean centered class average of pretest scores of the respective outcome measure. Treatment efficacy was tested at the class-level using dummy codes for experimental conditions (1 = SSIS-CIP, 0 = control). In addition to testing for treatment main effects, we examined if SSIS-CIP effects were moderated by pretest scores (both class- and student-levels) by adding product terms to the main effects model.[1]

Our centering approaches were chosen based on recommendations by Enders and Tofighi (2007). Specifically, all student-level covariates (except for the pretest of the outcome measure), including dummy demographic variables, were grand-mean centered to obtain adjusted treatment effects for these covariates. We modeled the pretest of the outcome in both the student- and class-levels because we were interested in the differential effect of pretest at these levels. Although group- and grand-mean centering the student-level pretest in this case would be equivalent (Enders & Tofighi, 2007), we chose to group-mean center the student-level and grand-mean center the class-level because the effect of pretest would be directly decomposed into between- and within-class levels (Raudenbush & Bryk, 2002). Moreover, group-mean centering student-level variables has been recommended in testing cross-level interactions (i.e., between treatment and student-level pretest) to minimize the risk of finding significant interactions that did not exist (e.g., Hofmann & Gavin, 1998).

We estimated multilevel models using the Mixed procedure of SAS (version 9.3) for teacher ratings of social skills and approaches to learning as well as for direct assessments of academic skills (mathematic and reading). We used the SAS Glimmix procedure for teacher ratings of problem behaviors and all direct observation data. Because problem behaviors were observed infrequently and observations consisted of frequency data that were highly skewed, we used Poisson distribution and log link for the Glimmix procedure.

For concerns of incorrect Type 1 error or false discovery rate in testing intervention effects across multiple outcome measures, we followed the What Works Clearinghouse (U.S. Department of Education, What Works Clearinghouse, 2016) recommendation and applied the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995) for teacher-rated social skills, problem behavior, and approaches to learning. Specifically, p values for treatment effects from the main effect models within the same outcome measure (e.g., teacher-rated social skills) were sorted from the lowest to the highest and compared with the corresponding critical p value. Critical p value was computed by multiplying alpha by the rank of the corresponding p value divided by the number of outcomes (i.e., the lowest p value was compared with .05*1/number of outcomes, the second lowest p value was compared to .05*2/number of outcomes, etc.). The highest p value that was less than or equal to the corresponding critical p value was the cutpoint, and findings with p values smaller than or equal to

[1] Consistent with our original proposal to the Institute of Education Sciences, we also examined if SSIS-CIP effects were moderated by student gender, minority status, and special education. Results of these analyses are reported in the supplement to this article.

Table 4
Class-Level Means (SDs) by Measures, Time, and Treatment Condition

	Pretest		Posttest	
Variable	SSIS-CIP	Control	SSIS-CIP	Control
Social skills				
Social skills composite	1.93 (.27)	2.19 (.28)	2.22 (.30)	2.31 (.27)
Communication	2.02 (.34)	2.31 (.28)	2.30 (.34)	2.45 (.31)
Cooperation	1.92 (.33)	2.14 (.32)	2.15 (.31)	2.22 (.29)
Assertion	1.71 (.31)	1.95 (.39)	2.02 (.42)	2.16 (.37)
Responsibility	2.01 (.32)	2.27 (.32)	2.25 (.28)	2.34 (.29)
Empathy	1.97 (.32)	2.27 (.31)	2.31 (.36)	2.39 (.33)
Social engagement	2.01 (.31)	2.25 (.32)	2.33 (.31)	2.42 (.32)
Self-control	1.91 (.31)	2.13 (.32)	2.15 (.36)	2.22 (.32)
Positive social[a]	.36 (.42)	.41 (.32)	.31 (.33)	.28 (.27)
Problem behaviors				
Externalizing	.60 (.29)	.49 (.29)	.55 (.27)	.49 (.28)
Bullying	.33 (.26)	.26 (.26)	.32 (.25)	.25 (.24)
Hyperactive-inattentive	.87 (.34)	.77 (.35)	.78 (.33)	.72 (.34)
Internalizing	.54 (.34)	.52 (.34)	.47 (.33)	.48 (.34)
Interference[a]	.16 (.13)	.28 (.32)	.19 (.23)	.33 (.43)
Approaches to learning				
Academic motivation	3.28 (.46)	3.44 (.57)	3.64 (.44)	3.61 (.54)
Academic engagement	3.54 (.48)	3.66 (.50)	3.94 (.47)	3.90 (.50)
Engaged time[a]	75.39 (9.91)	79.69 (8.16)	77.60 (10.67)	79.56 (10.06)
Academic skills				
Math scaled score	260.63 (84.57)	250.48 (76.78)	359.23 (65.23)	354.83 (64.28)
Reading scaled score	89.59 (42.42)	80.08 (31.62)	151.31 (53.32)	138.41 (46.25)

Note. SSIS-CIP = Social Skills Improvement System Classwide Intervention Program. SSIS-CIP $N = 28$; control $N = 29$.
[a] Direct observation data.

the cutpoint were declared statistically significant after the Benjamini-Hochberg correction.

Given the growing consensus that interpretation of study outcomes should not focus exclusively on statistical significance (e.g., Cumming, 2014; Durlak, 2009; Greenwald, Gonzalez, Harris, & Guthrie, 1996; Wasserstein & Lazar, 2016), we also estimated effect sizes of treatment as compared to the control condition based on the previously specified main effects models. Specifically, we computed the effect size as a standardized mean difference by dividing the adjusted (for pretest scores and other student- and class-level covariates) group mean difference by the unadjusted *pooled* within-group student-level *SD* of the pretest outcome measure. This effect size computation (i.e., using student-level *SD* to standardize the adjusted difference for Hedges' g) followed the guidelines of What Works Clearinghouse (U.S. Department of Education, What Works Clearinghouse, 2016) for results from Hierarchical Linear Model analyses in studies featuring cluster-level assignment. Pooled within-group *SD* of pretest scores was used because pretest scores were not affected by treatment. Moreover, 95% confidence intervals (CIs) were calculated for each effect size to provide insight regarding the precision of the estimate and range of possible effects. We also calculated an improvement index for each outcome variable following the U.S. Department of Education, What Works Clearinghouse (2016) guidelines to help readers better understand the practical impact of the SSIS-CIP intervention. An improvement index indicates the expected percentile rank improvement for an average student in the control group had the student received the intervention.

Results

Given the primary focus of the SSIS-CIP is to promote social skills in the classroom, our first hypothesis was that students in intervention classrooms would demonstrate improved social skills relative to their peers in control classrooms. Parameter estimates for the multilevel main effects models for the social skills domain are reported in Table 5. As shown in the table, two of the social skill domains (empathy and social engagement) were statistically significant ($p < .05$); however, the remaining teacher-rated social skills (communication, cooperation, assertion, responsibility, and self-control) and direct observation variable (positive social) did not demonstrate statistically significant differences (ps $> .05$). After applying the Benjamini-Hochberg (Benjamini & Hochberg, 1995) correction to control for false discovery rate, none of the observed differences met the adjusted threshold criterion for statistical significance. Tests of interactions between baseline level of social skills and intervention condition also did not demonstrate statistical significance (ps $> .05$) in any of the social skills domains (see Table 6). After controlling for pretest scores, gender, race, receipt of special education services, and receipt of supplemental (Title 1) services, SSIS-CIP participation yielded positive effect sizes (adjusted standardized difference) and at least five percentile rank improvement across all teacher-rated social skills subscales, with empathy and social engagement demonstrating the largest positive effects (see Table 7). The direct observation measure (positive social), however, yielded the smallest effect size and improvement index.

Table 5
Mixed Model Estimates (SEs) for SSIS-CIP Effect on Social Skills Outcomes

	Teacher rating								Direct observation
Predictor	Social skills composite	Commun.	Cooper.	Assertion	Responsib.	Empathy	Social engage.	Self-control	Positive social[a]
Intercept	2.13** (.05)	2.32** (.09)	2.06** (.06)	1.87** (.07)	2.16** (.06)	2.21** (.07)	2.25** (.06)	2.06** (.07)	−1.66** (.30)
Covariates									
Student-level pretest	70** (.03)	.21** (.07)	.72**(.05)	.56** (.03)	.65** (.06)	.36** (.06)	.37** (.06)	.61** (.05)	−.26 (.13)
Class-level pretest	−.003 (.11)	.46** (.13)	.64** (.11)	.70** (.11)	.50** (.12)	.52** (.11)	.50** (.12)	.59** (.12)	.78** (.24)
Social skills pretest	—	.45** (.08)	.01(.07)	.07 (.04)	.10 (.08)	.36** (.07)	.35** (.07)	.09 (.06)	.26 (.24)
Gender	−.06* (.02)	−.04 (.03)	−.11**(.03)	−.05 (.03)	−.06* (.03)	−.06 (.03)	−.01 (.03)	−.06 (.03)	.21 (.22)
White	−.001 (.03)	−.01 (.04)	−.01(.04)	.04 (.04)	−.003 (.04)	−.005 (.04)	−.01 (.04)	−.002 (.04)	.19 (.27)
Supp. services	−.01 (.03)	.02 (.03)	−.03(.04)	−.08* (.03)	−.02 (.04)	.03 (.04)	−.02 (.03)	.03 (.04)	.05 (.27)
Special education	.05 (.05)	.01 (.06)	.13(.07)	−.03 (.06)	.11 (.06)	.07 (.07)	−.01 (.06)	.07 (.07)	.42 (.41)
Treatment effect									
SSIS-CIP	.09 (.05)	.09 (.07)	.09 (.05)	.08 (.07)	.08 (.06)	.18 (.07)	.12 (.06)	.09 (.06)	.13 (.23)
	$p = .098$	$p = .199$	$p = .099$	$p = .247$	$p = .147$	$p = .012$	$p = .045$	$p = .152$	$p = .568$
Random effects									
Intercept variance	.03** (.01)	.04** (.01)	.02** (.01)	.04** (.01)	.03** (.01)	.04** (.01)	.03** (.01)	.03** (.01)	.07 (.13)
Residual variance	.09** (.005)	.12** (.01)	.17** (.01)	.12** (.01)	.15** (.01)	.16** (.01)	.13** (.01)	.16** (.01)	—

Note. School indicators are included in the model but not reported. SSIS-CIP = Social Skills Improvement System Classwide Intervention Program; Commun. = communication; Cooper. = cooperation; Responsib. = responsibility; Social engage. = social engagement; Supp. services = supplemental services.
[a] Outcome variable is on log scale.
$^{*}p < .05$. $^{**}p < .01$.

Our second hypothesis was that children exposed to the SSIS-CIP demonstrate fewer problem behaviors than their peers in control classrooms. Parameter estimates for the multilevel main effects models for problem behavior outcomes are reported in Table 8, and none of the differences between conditions on any of the problem behavior subscales were statistically significant (*ps* > .05). Similarly, tests of interactions between baseline level of problem behaviors and intervention condition also were not statistically significant (see Table 9). Effect size estimates for the problem behaviors subscales were consistently close to 0 with improvement indexes <3 (see Table 7).

Table 6
Mixed Model Estimates (SEs) for SSIS-CIP and Pretest Interaction on Social Skills Outcomes

	Teacher rating								Direct observation
Predictor	Social skills composite	Commun.	Cooper.	Assertion	Responsib.	Empathy	Social engage.	Self-control	Positive social[a]
Intercept	2.14** (.06)	2.33** (.08)	2.06** (.06)	1.89** (.07)	2.16** (.06)	2.20** (.07)	2.25** (.07)	2.07** (.07)	−1.70** (.30)
Covariates									
Student-level pretest	.71** (.04)	.20* (.08)	.71** (.06)	.53** (.04)	.65** (.07)	.39** (.07)	.38** (.07)	.63** (.06)	−.32 (.26)
Class-level pretest	−.08 (.14)	.43* (.17)	.63** (.13)	.60** (.13)	.49** (.14)	.53** (.15)	.50** (.14)	.53** (.15)	1.21* (.55)
Social skills pretest	—	.71** (.04)	.01 (.07)	.07 (.04)	.10 (.08)	.36** (.07)	.35** (.07)	.10 (.06)	.31 (.25)
Gender	−.06* (.02)	−.03 (.03)	−.11** (.03)	−.05 (.03)	−.06* (.03)	−.06* (.03)	−.02 (.03)	−.06 (.03)	.21 (.22)
White	−.001 (.03)	−.01 (.04)	−.01 (.04)	.05 (.04)	−.003 (.04)	−.004 (.04)	−.01 (.04)	−.002 (.04)	.22 (.28)
Supp. services	−.01 (.03)	.02 (.03)	−.04 (.04)	−.08* (.03)	−.02 (.04)	.03 (.04)	−.02 (.03)	.03 (.04)	.05 (.27)
Special education	.05 (.05)	.01 (.06)	.13 (.07)	−.03 (.06)	.10 (.07)	.07 (.07)	−.01 (.06)	.08 (.07)	.41 (.42)
Treatment effect									
SSIS-CIP	.09 (.05)	.09 (.07)	.09 (.06)	.09 (.07)	.08 (.06)	.18* (.07)	.12* (.06)	.09 (.06)	.21 (.25)
Interaction effect									
SSIS-CIP*Student- level	−.02 (.05)	.01 (.05)	.02 (.05)	.06 (.05)	.001 (.05)	−.05 (.06)	−.03 (.06)	−.05 (.06)	.08 (.30)
pretest	$p = .773$	$p = .806$	$p = .685$	$p = .281$	$p = .982$	$p = .382$	$p = .589$	$p = .424$	$p = .781$
SSIS-CIP*Class- level	.18 (.19)	.05 (.22)	.02 (.17)	.29 (.20)	.02 (.18)	−.03 (.20)	.003 (.19)	.12 (.19)	−.53 (.62)
pretest	$p = .349$	$p = .816$	$p = .883$	$p = .156$	$p = .893$	$p = .865$	$p = .987$	$p = .541$	$p = .399$
Random effects									
Intercept variance	.03** (.01)	.04** (.01)	.02** (.01)	.04** (.01)	.03** (.01)	.04** (.01)	.03** (.01)	.03** (.01)	.08 (.14)
Residual variance	.09** (.005)	.12** (.01)	.17** (.01)	.12** (.01)	.15** (.01)	.16** (.01)	.13** (.01)	.16** (.01)	—

Note. School indicators are included in the model but not reported. SS1S-CIP = Social Skills Improvement System Classwide Intervention Program; Commun. = communication; Cooper. = cooperation; Responsib. = responsibility; Social engage. = social engagement; Supp. services = supplemental services.
[a] Outcome variable is on log scale.
$^{*}p < .05$. $^{**}p < .01$.

Table 7
Standardized Group Differences, 95% Confidence Intervals, and Improvement Indices

Variable	Effect size[a]	95% Confidence interval	Improvement index (%)
Social skills			
Social skills composite	.18	[.03, .33]	7.14
Communication	.16	[.01, .30]	6.36
Cooperation	.14	[−.01, .29]	5.57
Assertion	.13	[−.02, .28]	5.17
Responsibility	.14	[.00, .29]	5.57
Empathy	.31	[.16, .46]	12.17
Social engagement	.21	[.06, .36]	8.32
Self-control	.15	[.00, .3]	5.96
Positive social[b]	.05	[−.17, .27]	1.99
Problem behaviors			
Externalizing	−.04	[−.19, .11]	−1.60
Bullying	.01	[−.14, .16]	.40
Hyperactive-inattentive	−.03	[−.18, .12]	−1.20
Internalizing	−.03	[−.18, .12]	−1.20
Interference[b]	−.07	[−.29, .15]	−2.79
Approaches to learning			
Academic motivation	.17	[.02, .32]	6.75
Academic engagement	.17	[.02, .32]	6.75
Engaged time[b]	.13	[−.09, .35]	5.17
Academic skills			
Math scaled score	.04	[−.11, .19]	1.60
Reading scaled score	.07	[−.08, .22]	2.79

[a] Standardized difference adjusted for pretest and other student- and class-level covariates. [b] Direct observation data.

The third hypothesis was that children exposed to the SSIS-CIP demonstrate improved approaches to learning (i.e., academic motivation and engagement), and parameter estimates for the multilevel main effects models for the approaches to learning domain are reported in Table 10. Teacher-ratings of academic motivation and engagement were statistically significant (*ps* $< .05$), and remained statistically significant after the Benjamini-Hochberg correction. Neither direct observation of engaged time (see Table 10) nor tests of interactions between pretest level of skills 'and intervention (see Table 11) were statistically significant. Effect size estimates indicated positive effects of SSIS-CIP exposure on academic motivation and engagement with improvement indexes >5 (see Table 7). Finally, our fourth hypothesis was that SSIS-CIP students demonstrate improved academic skills (reading and mathematics) relative to their peers. Parameter estimates for direct (see Table 10) and interaction effects (see Table 11) were not statistically significant for either reading or mathematics. Effect size estimates were slightly positive but close to 0, with improvement indexes below 3 (see Table 7).

Discussion

The purpose of this project was to examine social, behavioral, and academic outcomes resulting from implementation of a universal social skills program in primary classrooms. Specifically, first grade classrooms within six schools were randomly assigned to treatment or business-as-usual control conditions. Teachers in the classrooms assigned to treatment were formally trained and implemented the SSIS-CIP over a 12-week period. Outcomes were assessed via teacher ratings and direct observations of classroom behavior as well as computer-adaptive tests of reading and mathematics. Multiple indices (*p* values, effect sizes, CIs, and improvement indexes) were examined to draw conclusions about hypotheses.

SSIS-CIP Outcomes for Students in First Grade

With regard to the first hypothesis, none of the observed differences across the social skills subscales met a $p < .05$ threshold after correcting for false discovery rate using the Benjamini-Hochberg procedure. Effect sizes, CIs, and improvement indexes, though, suggest that the SSIS-CIP generally has small positive

Table 8
Multilevel Model Estimates (SEs) for SSIS-CIP Effect on Problem Behaviors Outcomes

	Teacher rating				Direct observation
Predictor	Externalizing	Bullying	Hyperactive-inattentive	Internalizing	Interference
Intercept	−.94** (.12)	−1.76** (.19)	−.52** (.10)	−1.05** (.14)	−1.97** (.39)
Covariates					
Student-level pretest	.87** (.13)	.82** (.15)	.73** (.10)	.68** (.15)	.36 (.23)
Class-level pretest	1.27** (.21)	1.73** (.32)	.96** (.17)	1.61** (.19)	2.47** (.51)
Social skills pretest	−.21 (.17)	−.63** (.20)	−.11 (.14)	−.40* (.16)	−.53* (.27)
Gender	.14 (.12)	−.07 (.16)	.20* (.10)	.05 (.12)	.70** (.26)
White	.08 (.13)	.09 (.18)	.08 (.11)	.003 (.14)	.17 (.29)
Supplemental services	.04 (.12)	−.04 (.17)	.07 (.10)	.01 (.12)	−.06 (.29)
Special education	−.14 (.21)	.04 (.30)	−.13 (.18)	−.01 (.22)	−.75 (.53)
Treatment effect					
SSIS-CIP	−.05 (.11) $p = .652$	.03 (.17) $p = .849$	−.03 (.10) $p = .751$	−.04 (.12) $p = .722$	−.18 (.32) $p = .578$
Random effects					
Intercept variance	<.0001	.05 (.070)	<.0001	.02 (.03)	.18 (.15)

Note. SSIS-CIP = Social Skills Improvement System Classwide Intervention Program. Outcome variables are on log scale. School indicators are included in the model but not reported.
* $p < .05$. ** $p < .01$.

Table 9
Multilevel Model Estimates (SEs) for SSIS-CIP and Pretest Interaction Effect on Problem Behaviors Outcomes

	Teacher rating			Direct observation	
Predictor	Externalizing	Bullying	Hyperactive-inattentive	Internalizing	Interference
Intercept	$-.97^{**}$ (.13)	-1.74^{**} (.19)	$-.55^{**}$ (.11)	-1.01^{**} (.14)	-1.91^{**} (.39)
Covariates					
Student-level pretest	1.00^{**} (.17)	$.79^{**}$ (.20)	$.82^{**}$ (.12)	$.66^{**}$ (.21)	.36 (.25)
Class-level pretest	1.24^{**} (.31)	1.35^{**} (.50)	$.98^{**}$ (.24)	1.38^{**} (.28)	2.28 (.53)
Social skills pretest	−.21 (.17)	$-.65^{**}$ (.20)	−.12 (.14)	$-.43^{**}$ (.16)	−.52 (.27)
Gender	.16 (.12)	−.09 (.16)	$.20^{*}$ (.10)	.05 (.12)	.70 (.26)
White	.08 (.13)	.09 (.18)	.10 (.11)	−.001 (.14)	.19 (.29)
Supplemental services	.02 (.12)	−.04 (.17)	.06 (.10)	.01 (.12)	−.06 (.29)
Special education	−.19 (.22)	.04 (.30)	−.16 (.18)	−.003 (.22)	−.77 (.54)
Treatment effect					
SSIS-CIP	.01 (.13)	−.05 (.20)	.02 (.11)	−.12 (.15)	−.23 (.32)
Interaction effect					
SSIS-CIP*Student-level pretest	−.20 (.17)	.05 (.24)	−.14 (.13)	.01 (.23)	.07 (.67)
	$p = .236$	$p = .841$	$p = .288$	$p = .957$	$p = .916$
SSIS-CIP*Class-level pretest	.03 (.39)	.61 (.62)	−.03 (.29)	.47 (.41)	1.83 (1.69)
	$p = .941$	$p = .324$	$p = .905$	$p = .252$	$p = .281$
Random effects					
Intercept variance	<.0001	.05 (.07)	<.0001	.01 (.03)	.17 (.15)

Note. SSIS-CIP = Social Skills Improvement System Classwide Intervention Program. Outcome variables are on log scale. School indicators are included in the model but not reported.
$^{*}p < .05.$ $^{**}p < .01.$

effects on the social skills of first grade students. Specifically, communication, cooperation, assertion, responsibility, and self-control all yielded effect sizes between .13–.16 with CIs ranging from approximately 0–.30 and improvement indexes in the 5–6 range. Social engagement and empathy had somewhat larger effect size estimates, higher CI bounds, and improvement indexes. (Conversely, direct observation of positive social behavior demonstrated a smaller effect size and improvement index.) The observed pattern of findings in first grade is similar to the pattern observed in the second grade trial of the SSIS-CIP (DiPerna et al., 2015). It is important to note, though, that effect sizes and improvement indexes for the first grade sample are consistently smaller (by approximately half) than those observed for second grade. The lone exception is empathy, where the effect sizes and corresponding CIs are almost identical (and more moderate in magnitude) across both trials.

Although the adjusted standardized group differences for social skills would be classified as small under Cohen's, 1988 guidelines, methodologists (Durlak, 2009; Ferguson, 2009) caution against rigid application of such guidelines and encourage interpretation

Table 10
Multilevel Model Estimates (SEs) for SSIS-CIP Effect on Approaches to Learning and Academic Skills Outcomes

	Teacher rating		Direct observation	Direct assessment	
Predictor	Academic motivation	Academic engagement	Engaged time[a]	Math	Reading
Intercept	3.38^{**} (.08)	3.64^{**} (.09)	4.35^{**} (.03)	365.02^{**} (12.01)	158.17^{**} (6.06)
Covariates					
Student-level pretest	$.69^{**}$ (.04)	$.57^{**}$ (.03)	$.002^{*}$ (.001)	$.45^{**}$ (.03)	$.84^{**}$ (.03)
Class-level pretest	$.65^{**}$ (.09)	$.61^{**}$ (.09)	$.01^{**}$ (.002)	$.57^{**}$ (.09)	$.98^{**}$ (.09)
Social skills pretest	.10 (.08)	$.16^{*}$ (.06)	$.09^{**}$ (.02)	45.17^{**} (7.94)	19.67^{**} (4.94)
Gender	$-.15^{**}$ (.05)	−.08 (.04)	$-.04^{**}$ (.01)	15.19^{*} (7.13)	−8.59 (4.54)
White	−.03 (.06)	.08 (.06)	$.04^{*}$ (.02)	14.66 (9.25)	13.71^{*} (5.88)
Supplemental services	$-.19^{**}$ (.06)	−.10 (.05)	.01 (.02)	-30.88^{**} (8.50)	-28.03^{**} (5.28)
Special education	−.05 (.10)	−.06 (.09)	−.01 (.03)	−21.36 (14.95)	−5.03 (9.52)
Treatment effect					
SSIS-CIP	.18 (.08)	.17 (.08)	.02 (.03)	5.61 (10.04)	5.87 (5.49)
	$p = .025$	$p = .035$	$p = .389$	$p = .579$	$p = .290$
Random effects					
Intercept variance	$.04^{**}$ (.01)	$.06^{**}$ (.02)	.01 (.002)	649.75^{**} (265.82)	1180.63 (75.92)
Residual variance	$.35^{**}$ (.02)	$.29^{**}$ (.02)	—	7820.35^{**} (438.76)	3212.35^{**} (179.73)

Note. SSIS-CIP = Social Skills Improvement System Classwide Intervention Program. School indicators are included in the model but not reported.
[a] Outcome variable is on log scale.
$^{*}p < .05.$ $^{**}p < .01.$

Table 11

Multilevel Model Estimates (SEs) for SSIS-CIP and Pretest Interaction Effect on Approaches to Learning and Academic Skills Outcomes

	Teacher rating		Direct observation	Direct assessment	
Predictor	Academic motivation	Academic engagement	Engaged time^{a}	Math	Reading
Intercept	3.38** (.08)	3.64** (.09)	77.23** (2.43)	365.22** (12.16)	158.40** (6.05)
Covariates					
Student-level pretest	.72** (.05)	.58** (.04)	.11 (.08)	.48** (.04)	.83** (.04)
Class-level pretest	.65** (.10)	.55** (.12)	.74** (.18)	.58** (.12)	1.08** (.15)
Social skills pretest	.10 (.08)	.16* (.06)	6.44** (1.49)	45.46** (7.95)	19.58** (4.95)
Gender	−.16** (.05)	−.08 (.04)	−2.94* (1.23)	14.48* (7.15)	−8.37 (4.56)
White	−.04 (.06)	.08 (.06)	3.00 (1.60)	14.43 (9.25)	13.48* (5.90)
Supplemental services	−.19** (.06)	−.10 (.05)	.37 (1.57)	−30.66** (8.50)	−28.16** (5.29)
Special education	−.03 (.10)	−.05 (.09)	−.33 (2.68)	−20.66 (14.97)	−5.07 (9.56)
Treatment effect					
SSIS-CIP	.18* (.08)	.17* (.08)	1.97 (2.08)	5.71 (10.14)	5.83 (5.48)
Interaction effect					
SSIS-CIP*Student-level pretest	−.06 (.05)	−.01 (.05)	.07 (.11)	−.07 (.05)	.02 (.06)
	p = .204	p = .805	p = .542	p = .186	p = .791
SSIS-CIP*Class-level pretest	.01 (.16)	.14 (.17)	−.20 (.24)	−.02 (.13)	−.14 (.16)
	p = .938	p = .434	p = .420	p = .847	p = .390
Random effects					
Intercept variance	.05** (.02)	.06** (.02)	34.96** (11.46)	675.77** (275.00)	115.86 (77.10)
Residual variance	.35** (.02)	.29** (.02)	105.52** (9.38)	7811.71** (438.74)	3219.67** (180.41)

Note. SSIS-CIP = Social Skills Improvement System Classwide Intervention Program. School indicators are included in the model but not reported.
^{a} Outcome variable is on log scale.
$^{*}p < .05$. $^{**}p < .01$.

relative to methodology and previous studies of similar interventions. Relative to randomized trials of other universal SEL programs, SSIS-CIP effect sizes based on teacher ratings of students' social skills are similar in magnitude to those reported for Second Step (g = −.016–.126; Low et al., 2015). Direct comparison with IYCD is difficult, as randomized trials evaluating its efficacy have included implementation of other Incredible Years interventions in conjunction with IYCD and outcome measures have not included teacher ratings of social skills (Baker-Henningham et al., 2009; Webster-Stratton et al., 2008). The 95% CI ranges for all of the SSIS-CIP social skills outcomes overlap with the mean effect sizes for student prosocial behavior reported by two recent meta-analyses (Durlak et al., 2011; January et al., 2011).

With regard to problem behaviors, results were inconsistent with our hypothesis that exposure to the SSIS-CIP would yield reductions in these behaviors. All tests of significance in this domain were nonsignificant with effect sizes and improvement indexes close to 0, and 95% effect size CIs that extended in both the positive and negative directions. Similar to the social skills domain, the problem behavior effect sizes observed in the first grade trial are smaller than those observed in the second grade trial that ranged from −.08 (improvement index = 3.19) for bullying to −.24 (improvement index = 9.48) for internalizing behavior (DiPerna et al., 2015). Relative to previous research, the SSIS-CIP first grade effect sizes are lower than those reported for problem behavior in the Durlak et al. (2011) meta-analysis (g = .22–.24). The studies in their review, however, included a much broader range of behavioral (e.g., noncompliance, aggression, and school suspensions) and emotional problems (e.g., stress, depression, and anxiety) than the measures in the current study. Results from the current study are more similar to those from the recent efficacy trial for Second Step (Low et al., 2015) in which specific problem behavior outcomes were measured through teacher ratings and direct observations. The Second Step effect size for conduct problems (g = −.04) was equivalent to SSIS-CIP effect size for externalizing behaviors, and although the other Second Step problem behavior outcomes effect sizes were larger ($-.109 < g < -.067$), they still fell within the 95% CI ranges for similar problem behavior outcomes in the current study. Similar to the current study, Webster-Stratton et al. (2008) reported a small mean effect size (.032) for IYCD on conduct problems measured via direct observation.

Our third hypothesis was that SSIS-CIP improves young students' approaches to learning (academic motivation and engagement), and results were consistent with this hypothesis. Teacher ratings of academic motivation and engagement remained statistically significant after the Benjamini-Hochberg correction and yielded small effect sizes with positive CIs and improvement indexes. As with the social skills domain, the first grade effect sizes and improvement indexes based on teacher ratings of academic motivation and engagement were approximately half the magnitude of those in the second grade trial (DiPerna et al., 2016). In contrast to all other findings, the effect size for direct observation of engagement in instruction was larger in the current study (.13) than the second grade trial (.03). Unfortunately, few CRTs of other universal SEL interventions (Durlak et al., 2011; January et al., 2011; Nelson et al., 2003) have assessed these intermediary variables hypothesized to link classroom behavior to academic outcomes. Low et al. (2015) did observe a positive effect of Second Step on a "social-emotional skills for learning" variable, which appears to overlap with the skills and attitudes of the academic motivation and engagement variables in the current study. The observed Second Step effect size (g = .114) was slightly smaller in magnitude than (but still within the 95% CI of) the effect sizes observed in the current study.

The final hypothesis was that students exposed to the SSIS-CIP curriculum would demonstrate improved academic skills relative to their peers in nonimplementing classrooms. Results, however, did not support this hypothesis as all tests were nonsignificant with effect sizes and improvement indexes close to 0. In addition, the 95% CIs for both reading and mathematics effect sizes ranged from positive to negative. These findings were consistent with the academic skills effect sizes reported in the second grade trial (DiPerna et al., 2016); however, they are smaller than the mean effect sizes reported in the Durlak et al. (2011) meta-analysis (g = .27). One possible explanation for these differences is that Durlak et al.'s review included studies that used school grades as well as standardized tests as outcome measures. Additionally, the reviewed studies focused on a wide range of programs, including multicomponent programs that supplemented teacher-facilitated programs with parent or school-wide initiatives. It is unknown if some of these components may have had an academic enrichment focus, as has been the case in some studies of SEL programs (e.g., Bradshaw et al., 2009). Finally, given that SEL and approaches to learning may set the foundation for development of school readiness and academic achievement (Blair & Raver, 2015), academic outcomes may be more apparent in research evaluating the long-term effects of a universal social-emotional program (e.g., Nelson et al., 2003).

In addition to testing for main effects of SSIS-CIP implementation in each of the four outcome domains (social skills, problem behaviors, approaches to learning, and academic skills), we tested interactions between initial skill level at pretest (at both the student and classroom level) and intervention condition within each analysis. In the second Grade SSIS-CIP trial (DiPerna et al., 2015, 2016) as well as studies of IYCD (Webster-Stratton et al., 2008) and Second Step effectiveness (Low et al., 2015), students with lower levels of initial skills demonstrated larger positive effects relative to their peers with higher levels of initial skills. In the current study, however, there were no statistically significant initial skill level by treatment interactions within any of the four skill domains for first grade students. Although this finding must be replicated, it suggests that, though the SSIS-CIP effects may be smaller in magnitude for younger students in the primary grades, they also may be more universally distributed throughout the classroom.

Limitations and Directions for Future Research

There are several limitations to the current study that also provide directions for future research. First, though the study included a sufficient number of classrooms and students to detect small effects, these participants were drawn from a limited number of schools across two school districts (rural and small urban). As such, replication of the current findings with an additional sample of first grade classrooms and schools is necessary. In addition, although the different versions of the SSIS-CIP curriculum for the intermediate (Grades 3–5) and preschool levels are similar to the version tested in this study, efficacy trials focused on those versions are necessary to determine if the impact of the curriculum is similar across developmental levels.

Beyond these design and replication considerations, it also is important to note that the student observation system used in the current study (CLOCK) focused on molar (broader) classes of behavior to make it feasible for staff to observe student behavior across all three outcome domains of interest (social skills, problem behaviors, and academic engagement). Despite using this approach and completing multiple direct observations for each student, the means for these variables were still low. In addition, although the effect size CIs overlapped between CLOCK scores and corresponding scores from the teacher rating scales, the strongest evidence for the social skills and approaches to learning predictions was based on teacher report via measures well-aligned with the SSIS-CIP program. Given teachers are not blind to treatment condition, future studies of the SSIS-CIP may benefit from using an observation protocol that assesses the specific social skills targeted by the SSIS-CIP instructional units (e.g., empathy, assertion, and self-control). Finally, the current study focused on immediate outcomes resulting from the SSIS-CIP. Studies examining follow-up data, as well as resources required for implementation, are necessary to better understand the benefits and costs associated with this universal program.

Conclusion

Results from the current study suggest that exposure to the SSIS-CIP curriculum has small positive effects on first graders' social skills and approaches to learning. Although the observed effect sizes in these domains were consistent with those reported for other classwide programs (e.g., January et al., 2011; Low et al., 2015), they were approximately half the magnitude of those observed when the SSIS-CIP was implemented with second-grade students (DiPerna et al., 2015, 2016). Effects on problem behaviors were negligible and lower than the effects observed both in the second Grade SSIS-CIP trial (DiPerna et al., 2015) and Durlak et al.'s (2011) meta-analyses of other universal SEL interventions. Academic skill effect sizes also were negligible, which was consistent with outcomes of the second grade trial but smaller than other universal SEL interventions (Durlak et al., 2011).

The pattern of observed findings across first and second grade suggests that the SSIS-CIP Early Elementary version yields positive effects in the social skills and approaches to learning domains with the effects being larger when implemented in second grade. It is important to note, though, that the 95% CIs for the effect sizes demonstrate overlap across all domains. Thus, it is possible that the observed differences in magnitude between first and second grade are not true differences but because of random variation across trials. As such, additional trials are necessary to determine if the SSIS-CIP effects replicate with new samples of first and second-grade students (e.g., Makel & Plucker, 2014). SSIS-CIP implementation appears to have negligible immediate effects on first grade students' problem behaviors and academic skills, though again given the 95% CIs around the effect sizes in the current study, replication is necessary to confirm these findings. If the pattern of findings across the first- and second-grade trials are replicated in future studies, educators and administrators contemplating adoption of the SSIS-CIP should consider prioritizing second grade for implementation of the program. In addition, researchers studying other universal SEL programs for young children should test for potential developmental differences in program effectiveness.

References

Ashdown, D. M., & Bernard, M. E. (2012). Can explicit instruction in social and emotional learning skills benefit the social-emotional development, well-being, and academic achievement of young children? *Early Childhood Education, 39,* 397–405. http://dx.doi.org/10.1007/s10643-011-0481-x

Baker-Henningham, H., Walker, S., Powell, C., & Gardner, J. M. (2009). A pilot study of the Incredible Years Teacher Training programme and a curriculum unit on social and emotional skills in community preschools in Jamaica. *Child: Care, Health and Development, 35,* 624–631. http://dx.doi.org/10.1111/j.1365-2214.2009.00964.x

Barrish, H. H., Saunders, M., & Wolf, M. M. (1969). Good behavior game: Effects of individual contingencies for group consequences on disruptive behavior in a classroom. *Journal of Applied Behavior Analysis, 2,* 119–124. http://dx.doi.org/10.1901/jaba.1969.2-119

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B. Methodological, 57,* 289–300.

Bierman, K. L., Domitrovich, C. E., Nix, R. L., Gest, S. D., Welsh, J. A., Greenberg, M. T., . . . Gill, S. (2008). Promoting academic and soeial-emotional school readiness: The head start REDI program. *Child Development, 79,* 1802–1817. http://dx.doi.org/10.1111/j.1467-8624.2008.01227.x

Blair, C., & Raver, C. C. (2015). School readiness and self-regulation: A developmental psychobiological approach. *Annual Review of Psychology, 66,* 711–731. http://dx.doi.org/10.1146/annurev-psych-010814-015221

Bradshaw, C. P., Zmuda, J. H., Kellam, S. G., & Ialongo, N. S. (2009). Longitudinal impact of two universal preventive interventions in first grade on educational outcomes in high school. *Journal of Educational Psychology, 101,* 926–937. http://dx.doi.org/10.1037/a0016586

Bub, K. L. (2009). Testing the effects of classroom supports on children's social and behavioral skills at key transition points using latent growth modeling. *Applied Developmental Science, 13,* 130–148. http://dx.doi.org/10.1080/10888690903041527

Cohen, J. (1988). *Statistical power for analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Collaborative for Academic, Social, and Emotional Learning. (2016, October 5). *Approaches to social-emotional learning.* Retrieved from http://www.casel.org/what-is-sel/approaches/

Committee for Children. (1992). *Second Step: A violence prevention curriculum, grades 1–3* (2nd ed.). Seattle, WA: Author.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25,* 7–29. http://dx.doi.org/10.1177/0956797613504966

Denham, S. A., Blair, K. A., DeMulder, E., Levitas, J., Sawyer, K., Auerbach-Major, S., & Queenan, P. (2003). Preschool emotional competence: Pathway to social competence? *Child Development, 74,* 238–256. http://dx.doi.org/10.1111/1467-8624.00533

Denham, S. A., Way, E., Kalb, S. C., Warren-Khot, H. K., & Bassett, H. H. (2013). Preschoolers' social information processing and early school success: The challenging situations task. *British Journal of Developmental Psychology, 31,* 180–197. http://dx.doi.org/10.1111/j.2044-835X.2012.02085.x

DiPerna, J. C., & Elliott, S. N. (2000). *Academic Competence Evaluation Scales.* San Antonio, TX: The Psychological Corporation.

DiPerna, J. C., Lei, P., Bellinger, J., & Cheng, W. (2015). Efficacy of the Social Skills Improvement System Classwide Intervention Program (SSIS-CIP) primary version. *School Psychology Quarterly, 30,* 123–141. http://dx.doi.org/10.1037/spq0000079

DiPerna, J. C., Lei, P., Bellinger, J., & Cheng, W. (2016). Effects of a universal positive classroom behavior program on student learning. *Psychology in the Schools, 53,* 189–203. http://dx.doi.org/10.1002/pits.21891

DiPerna, J. C., Volpe, R. J., & Elliott, S. N. (2005). A model of academic enablers and mathematics achievement in the early grades. *Journal of School Psychology, 43,* 379–392. http://dx.doi.org/10.1016/j.jsp.2005.09.002

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology, 34,* 917–928. http://dx.doi.org/10.1093/jpepsy/jsp004

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82,* 405–432. http://dx.doi.org/10.1111/j.1467-8624.2010.01564.x

Durlak, J. A., Weissberg, R. P., & Pachan, M. (2010). A meta-analysis of after-school programs that seek to promote personal and social skills in children and adolescents. *American Journal of Community Psychology, 45,* 294–309. http://dx.doi.org/10.1007/s10464-010-9300-6

Early, D., Pianta, R., & Cox, M. (1999). Kindergarten teachers and classrooms: A transition context. *Early Education and Development, 10,* 25–46. http://dx.doi.org/10.1207/s15566935eed1001_3

Elliott, S. N., & Gresham, F. M. (2007). *Social Skills Improvement System: Classwide Intervention Program.* Minneapolis, MN: Pearson Assessments.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12,* 121–138. http://dx.doi.org/10.1037/1082-989X.12.2.121

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology, Research and Practice, 40,* 532–538. http://dx.doi.org/10.1037/a0015808

Flay, B. R., & Allred, C. G. (2003). Long-term effects of the Positive Action program. *American Journal of Health Behavior, 27*(Suppl. 1), S6–S21. http://dx.doi.org/10.5993/AJHB.27.1.s1.2

Flay, B. R., Allred, C. G., & Ordway, N. (2001). Effects of the Positive Action program on achievement and discipline: Two matched-control comparisons. *Prevention Science, 2,* 71–89. http://dx.doi.org/10.1023/A:1011591613728

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576. http://dx.doi.org/10.1146/annurev.psych.58.110405.085530

Greenberg, M. T., Domitrovich, C., & Bumbarger, B. (2001). The prevention of mental disorders in school-aged children: Current state of the field. *Prevention & Treatment, 4,* 1–62. http://dx.doi.org/10.1037/1522-3736.4.1.41a

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology, 33,* 175–183. http://dx.doi.org/10.1111/j.1469-8986.1996.tb02121.x

Gresham, F., & Elliott, S. N. (2008). *Social Skills Improvement System Rating Scales.* Minneapolis, MN: Pearson Assessments.

Grossman, D. C., Neckerman, H. J., Koepsell, T. D., Liu, P. Y., Asher, K. N., Beland, K., . . . Rivara, F. P. (1997). Effectiveness of a violence prevention curriculum among children in elementary school. A randomized controlled trial. *Journal of the American Medical Association, 277,* 1605–1611. http://dx.doi.org/10.1001/jama.1997.03540440039030

Hamre, K. B., Mashburn, A. J., Pianta, R. C., & LoCasale-Crouch, J. (2008). *Classroom Assessment Scoring System, Pre-K: Technical appendix.* Baltimore, MD: Brookes.

Heckman, J. J. (2006). A broader view of what education policy should be. In N. F. Watt, C. Ayoub, R. H. Bradley, J. E. Puma, & W. A. LeBoeuf (Eds.), *The crisis in youth mental health: Early intervention programs and policies* (Vol. 4, pp. 3–26). Westport, CT: Praeger.

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 24,* 623–641. http://dx.doi.org/10.1177/014920639802400504

Hughes, J. N., & Kwok, O. M. (2006). Classroom engagement mediates the effect of teacher-student support on elementary students' peer acceptance: A prospective analysis. *Journal of School Psychology, 43,* 465–480. http://dx.doi.org/10.1016/j.jsp.2005.10.001

January, A. M., Casey, R. J., & Paulson, D. (2011). A meta-analysis of classroom-wide interventions to build social skills: Do they work? *School Psychology Review, 40,* 242–256.

Linares, L. O., Rosbruch, N., Stern, M. B., Edwards, M. E., Walker, G., Abikoff, H. B., & Alvir, J. (2005). Developing cognitive-social-emotional competencies to enchance academic learning. *Psychology in the Schools, 42,* 405–417. http://dx.doi.org/10.1002/pits.20066

Low, S., Cook, C. R., Smolkowski, K., & Buntain-Ricklefs, J. (2015). Promoting social-emotional competence: An evaluation of the elementary version of Second Step®. *Journal of School Psychology, 53,* 463–477. http://dx.doi.org/10.1016/j.jsp.2015.09.002

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43,* 304–316. http://dx.doi.org/10.3102/0013189X14545513

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7,* 537–542. http://dx.doi.org/10.1177/1745691612460688

McClelland, M. M., Acock, A. C., & Morrison, F. J. (2006). The impact of kindergarten learning-related skills on academic trajectories at the end of elementary school. *Early Childhood Research Quarterly, 21,* 471–490. http://dx.doi.org/10.1016/j.ecresq.2006.09.003

Miles, S. B., & Stipek, D. (2006). Contemporaneous and longitudinal associations between social behavior and literacy achievement in a sample of low-income elementary school children. *Child Development, 77,* 103–117. http://dx.doi.org/10.1111/j.1467-8624.2006.00859.x

Nelson, G., Westhues, A., & MacLeod, J. (2003). A meta-analysis of longitudinal research on preschool prevention programs for children. *Prevention & Treatment, 6,* 1–35. http://dx.doi.org/10.1037/1522-3736.6.1.631a

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *The Classroom Assessment Scoring System Manual (K-3).* Baltimore, MD: Brookes.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2,* 173–185. http://dx.doi.org/10.1037/1082-989X.2.2.173

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Spybrook, J., Liu, X., & Congdon, R. (2005). Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" software [Computer program] (Version March 09, 2005).

Renaissance Learning. (2009). *STAR Math technical manual.* Wisconsin Rapids, WI: Renaissance Learning.

Renaissance Learning. (2010). *STAR Reading technical manual.* Wisconsin Rapids, WI: Renaissance Learning.

Shonkoff, J. P., & Philips, D. A. (Eds.). (2000). *From neurons to neighborhoods: The science of early childhood development.* Washington, DC: National Academy Press.

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2016, September). *What works clearinghouse: procedures and standards handbook* (Version 3.0). Retrieved from http://whatworks.ed.gov

Volpe, R. J., & DiPerna, J. C. (2010). *Cooperative learning observation code for kids.* Unpublished observation code.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70,* 129–133. http://dx.doi.org/10.1080/00031305.2016.1154108

Webster-Stratton, C., Jamila Reid, M., & Stoolmiller, M. (2008). Preventing conduct problems and improving school readiness: Evaluation of the Incredible Years Teacher and Child Training Programs in high-risk schools. *Journal of Child Psychology and Psychiatry, 49,* 471–488. http://dx.doi.org/10.1111/j.1469-7610.2007.01861.x

Webster-Stratton, C., & Reid, M. J. (2004). Strengthening social and emotional competence in young children-The foundation for early school readiness and success: Incredible Years Classroom Social Skills and Problem-Solving Curriculum. *Infants and Young Children, 17,* 96–113. http://dx.doi.org/10.1097/00001163-200404000-00002

Wentzel, K. R., & Asher, S. R. (1995). The academic lives of neglected, rejected, popular, and controversial children. *Child Development, 66,* 754–763. http://dx.doi.org/10.2307/1131948

Received October 19, 2015
Revision received January 5, 2017
Accepted January 9, 2017 ■

Journal of Educational Psychology
2018, Vol. 110, No. 1, 17–26

0022-0663/18/$12.00 http://dx.doi.org/10.1037/edu0000204

Child- and School-Level Predictors of Children's Bullying Behavior: A Multilevel Analysis in 648 Primary Schools

Elian Fink
University College London and University of Cambridge

Praveetha Patalay
University College London and University College London Institute of Education

Helen Sharpe
University College London and University of Edinburgh

Miranda Wolpert
Anna Freud Centre and University College London

A great deal of bullying behavior takes place at school, however, existing literature has predominantly focused on individual characteristics of children associated with bullying with less attention on school-level factors. The current study, comprising 23,215 children (51% boys) recruited from Year 4 or Year 5 (M = 9.06 years, SD = .56 years) from 648 primary schools in England, aimed to examine the independent and combined influence of child- and school-level predictors on bullying behavior in primary school. Children provided information on bullying behavior and school climate. Demographic characteristics of children were obtained from the National Pupil Database, and demographic characteristics of schools were drawn from EduBase. Multilevel logistic regression models showed that individual child gender, ethnicity, deprivation and special educational needs status all predicted bullying behavior. Of the school-level predictors, only overall school deprivation and school climate were predictive of bullying behavior once child-level predictors were taken into account. There was a significant interaction between child- and school-level deprivation; high-deprivation schools were a risk factor for bullying only for children that came from nondeprived backgrounds, whereas deprived children reported engaging in bullying behavior irrespective of school-level deprivation. Given the independent and combined role of child- and school-level factors for bullying behavior, the current study has implications for targeted school interventions to tackle bullying behavior, both in terms of identifying high-risk children and identifying high-risk schools.

Educational Impact and Implications Statement

The majority of bullying behavior takes place at school, however, the existing literature has mostly focused on child characteristics associated with bullying, and considerably less attention has been paid to the characteristics of children's schools that are associated with bullying. We analyze data from 23,215 children form 648 primary schools to identify both child and school characteristics that predict children's bullying behavior. A number of child characteristics were found to predict greater bullying behavior (such as being a boy and experiencing poverty). Of the school characteristics, aggregated poverty level and sense of school connectedness were associated with bullying behavior. Additionally, the statistical combination of child and school characteristics were also shown to predict bullying, such that children not experiencing poverty attending schools with high poverty levels were at particular risk of engaging in bullying behavior. The current study has important implications for the design and implementation of targeted school interventions to tackle bullying behavior, both in terms of identifying and targeting high-risk children and high-risk schools.

Keywords: bullying, school composition, school climate, multilevel analysis

This article was published Online First April 17, 2017.

Elian Fink, Evidence-Based Practice Unit, University College London; Faculty of Education, University of Cambridge; Praveetha Patalay, Evidence-Based Practice Unit, University College London; Centre for Longitudinal Studies, University College London Institute of Education; Helen Sharpe, Evidence-Based Practice Unit, University College London; Department of Clinical Psychology, University of Edinburgh; Miranda Wolpert, Evidence-Based Practice Unit, Anna Freud Centre and University College London.

We are grateful to all the participating staff and students for supporting this work and the other members and advisors of the research group who were involved in the wider Me and My School Project. We thank the English Department for Children, Schools and Families (now the Department for Education) for funding the wider study from which these data are drawn.

Correspondence concerning this article should be addressed to Elian Fink, Faculty of Education, University of Cambridge, 184 Hills Road, Cambridge, CB2 8PQ United Kingdom. E-mail: ef364@cam.ac.uk

Bullying at school is a significant problem and has a far-reaching negative influence on later psychosocial adjustment (e.g., Fisher et al., 2012; Glew, Fan, Katon, Rivara, & Kernic, 2005), educational attainment (e.g., Glew et al., 2005) and physical health (e.g., Takizawa, Maughan, & Arseneault, 2014). To aid in the understanding of the factors that predict bullying behavior, the extant literature has, for the most part, focused on individual characteristics of children, such as age, gender, externalising problems and social cognitions (Cook, Williams, Guerra, Kim, & Sadek, 2010). However, bullying behavior necessarily takes place in a social context and, bullying, by definition, is a relational process established over time (Salmivalli, 2010); thus, contextual factors, including school characteristics, are likely to play a key role in bullying behavior (Bradshaw, Sawyer, & O'Brennan, 2009; Cook et al., 2010). Understanding the school-level features predictive of bullying is especially pertinent given that the majority of bullying interventions stress the importance of changing the school environment (e.g., Olweus & Limber, 2010). However, relatively little systematic attention has been given to these features. In particular, very little focus has been placed on understanding the potential interaction between child- and school level predictors of bullying behavior. The current study examines both child- and school-level risk factors for bullying behavior in a large cohort of primary school-age children to better understand the predictors of bullying behavior in this population.

Understanding bullying in primary school is important, as these years are a critical developmental stage where children learn to establish and maintain peer relationships (e.g., Fink, Begeer, Hunt, & de Rosnay, 2014). Notably, studies with children and retrospective studies with adults have both shown that bullying experiences typically emerge during the primary school years, making this period a crucial time for understanding the child- and school factors associated with bullying behavior (Bowes et al., 2009; Wolke, Woods, Bloomfield, & Karstadt, 2000). Furthermore, bullying increases during childhood, peaking in early adolescence suggesting that prevention programs targeting children at the end of primary school prior to this peak may be the most effective at curbing this increase (Guerra et al., 2011). This suggests that understanding predictors of bullying behaviors in these earlier years may be critical to breaking a cycle of bullying that can perpetuate through adolescence (Smith, 2010).

School-Level Predictors

School is a key context for bullying behavior during childhood (Saarento, Kärnä, Hodges, & Salmivalli, 2013), and recent empirical research has begun to acknowledge the important role played by the child's specific school context for the prediction of bullying behavior (Bowes et al., 2009; Bradshaw et al., 2009; Khoury-Kassabri, Benbenishty, Astor, & Zeira, 2004; Saarento, Garandeau, & Salmivalli, 2015; Vervoort, Scholte, & Overbeek, 2010; Waasdorp, Pas, O'Brennan, & Bradshaw, 2011; Wolke, Woods, Stanford, & Schulz, 2001). This research has shown that between 1% and 7% of variability in bullying behavior is accounted for by the classroom or school group (Bradshaw et al., 2009; Kärnä et al., 2013; Kärnä et al., 2011), compared to around 20% for academic attainment and under 5% for mental health variables (Hale et al., 2014). A range of school-level characteristics may be considered in relation to bullying, including school composition, school climate, and the presence of bullying prevention and victim support strategies. The first and second of these aspects will be the focus of the current research, and their implications for interventions are discussed.

The most commonly studied school composition factors that predict bullying include the following: gender distribution, classroom or school size, proportion of ethnic minority students, and socioeconomic indices (see Kasen, Berenson, Cohen, & Johnson, 2004 for reviews; Saarento et al., 2015). However, results from these studies have been inconsistent. For example, while some research has shown that a greater proportion of boys within a school is associated with a greater bullying (e.g., Khoury-Kassabri et al., 2004) other studies have failed to find such an effect (e.g., Saarento et al., 2013). Similar discrepancies in the extant literature are also observed for classroom or school size, with different studies showing an advantage of either larger or smaller schools (or classrooms) for bullying (e.g., Bowes et al., 2009; Khoury-Kassabri et al., 2004; Saarento et al., 2013; Whitney & Smith, 1993). With respect to ethnic minority composition, some studies have found no association between proportion of students from ethnic minorities (e.g., Whitney & Smith, 1993), whereas others have found interactions between classroom composition of ethnic minorities and individual children's minority status as predictive of bullying (e.g., Vervoort et al., 2010). Finally, students from low socioeconomic status (SES) schools have been found to report greater bullying (e.g., Bradshaw et al., 2009; Whitney & Smith, 1993), while others have not found such an association (e.g., Ma, 2002). It is clear, therefore, from the extant literature, that the prediction of bullying from demographic school-level variables has produced varied findings. These differences may be due to diversity in bullying measurement across studies, differences across studies in controlling for child-level predictors and the age-group of the participants (primary or secondary school). In addition, few studies include a large number of schools, meaning that they likely lack power to detect between-school variations.

Contrary to the findings for school-level demographic factors, school climate, frequently operationalised as the extent to which students on average feel connected to their school and have positive perceptions of school (and their teachers), does appear to be consistently associated with bullying behavior (e.g., Bosworth, Espelage, & Simon, 1999; Bradshaw, Waasdorp, & Johnson, 2015; Guerra, Williams, & Sadek, 2011; Kasen et al., 2004). For example, in schools where victimization is a problem, children tend to report less positive perceptions of their school climate (Baker, 1998; Ma, 2002). Furthermore, children who report bullying others also report significantly more negative perceptions of and feel less connected to their school (Espelage & Swearer, 2003; Nansel et al., 2001). It is worth noting that the majority of this work has been conducted within North America, and it is not clear to what extent it will apply in other school contexts, such as those in Europe. These findings are, however, encouraging as they suggest that the malleable factor of school climate plays a role in the extent to which schools experience bullying, implying that providing schools with support to improve their school climate will also have a positive impact on bullying behavior.

An important distinction may be made between school composition and school climate as predictors of bullying behavior. School composition relates to the characteristics and mix of the students within a given school, and as such is a nonmalleable characteristic of the school. Although it is important to understand the impact of these variables on bullying behavior, they are not subject to direct intervention. However, studying the influence of school composition offers an insight into the environment that facilitates bullying behavior and where intervention can be targeted. School climate, commonly representing aggregated individual child perceptions of their school, conversely, is a dynamic aspect of schools that is malleable to intervention (Brault, Janosz, & Archambault, 2014), and interventions aiming to improve school-wide culture have been shown to decrease the incidence of behavior problems (Bradshaw, Waasdorp, & Leaf, 2012).

Child-Level Predictors

When considering school-level characteristics it is essential to understand if these school-level factors make an impact beyond the individuals that make up the school. Many child demographic factors have been implicated in bullying behavior. Most consistently gender has been shown to be associated with bullying behavior, with boys engaging in higher amounts of bullying behavior than girls (Bosworth et al., 1999; Cook et al., 2010). Children's socioeconomic status has also been found to be associated with bullying behavior although this result is not always consistent (see Wolke et al., 2001), highlighting the need to further explore this result and examine if it is child-level or school-level disadvantage that is more closely associated with bullying behavior. Similarly, ethnicity has also sometimes been shown to be associated with bullying behavior (e.g., Wolke et al., 2001) although again, this finding is not always consistent (e.g., Bosworth et al., 1999). As such, greater clarity is also needed on the individual demographic characteristics that put children at risk of exhibiting bullying behavior.

The Current Study

Given the limitations of the existing literature outlined above, the current study aimed to make two advances: (1) to examine school-level influences on bullying behavior in a large cohort of primary schools (648 schools), giving power to detect school-level effects, and (2) to explore the interplay between child- and school-level influences. Specifically, we explored the role of school size, school gender balance, proportion of children from minority ethnic groups, school deprivation and school climate as predictors of bullying behavior over and above individual child-level demographic characteristics (gender, deprivation, ethnicity, special educational needs status, English as an additional language and year group). Given the broad nature of school climate, we focused specifically on school supportiveness and school connectedness, which are most commonly examined within the bullying literature. By employing a multilevel modeling approach and including both child- and school-level factors simultaneously, the current study is able to assess the relative independent and combined impact of each for bullying behavior. Understanding the unique contribution of school-level factors that influence bullying has reaching implications for school-based interventions designed to curtail bullying in schools and promote a more positive school environment and can, furthermore, provide guidance for targeting bullying interventions to those schools that need them most.

Method

Participants.

Schools. A total of 648 primary schools participated in the current study. Schools were selected by their local authority to participate in a larger study examining child mental health across England (Wolpert et al., 2011). All schools were state-maintained (i.e., public schools) in England and provided an average of 35.80 participants per school (SD = 18.65). Details of school characteristics are presented in Table 1. Schools were drawn from 99 (of 351) local authorities across England, and the geographical spread of these local authorities was representative of the whole country. Out of the 648 schools in the study, one was a single-sex boys' schools and the remaining 647 were mixed sex schools. Eight schools were focused on students with special educational needs.

Children. The study included 23,215 participants (51% boys) from Year 4 and Year 5 (M_{age} = 9.06 years, SD = .56 years). All children in Year 4 or Year 5 at the selected schools were invited to participate and consent was sought from parents beforehand by post and children provided assent prior to completing measures.

With respect to the ethnicity of participants in the current study, 75% were recorded as White, with the remainder being Asian (12%), Black (7%), mixed (4%), or other ethnic groups/unclassified (2%). Comparing these proportions to the overall proportions of children from black and ethnic minority groups (BME) for the whole school (i.e., all year levels of participating schools; see Table 1) shows that the subsample of Year 4 and 5 students in the current study largely mirror the overall school composition. Furthermore, when comparing the proportion of BME students in our current sample to all students attending primary schools across England (Department for Education, 2009), again, shows that our sample is representative of the total primary school population.

Socioeconomic status was based on children's eligibility for free school meals (FSM; Hobbs & Vignoles, 2010). Of participating children, 24% were eligible for FSM, which is higher than the national average of 16% (Department for Education, 2009). Finally, 20% of the sample knew English as an additional language (EAL) and 28% were identified as having any special educational needs (SEN).

Procedure. Data were collected from three sources. First, children completed self-report measures using a secure online system during their usual school day. A description of the full battery of measures and study design is reported elsewhere (Wolpert et al., 2011). Teachers facilitated the completion of the survey

Table 1
School Characteristics (n = 648)

Characteristic	*M (SD)*	Range
Size (number of pupils)	304.00 (135.85)	29–1212
Gender (% girls)	48.45 (4.57)	.00–60.00
Deprivation (% FSM)	18.73 (12.93)	.00–87.50
Ethnicity (% BME)	22.67 (28.04)	.00–100
SEN %	29.71 (15.37)	.00–100
EAL %	17.12 (25.35)	.00–100
School climate	11.77 (.83)	7.08–13.81

Note. FSM = free school meals; BME = Black and minority ethnic; SEN = special educational needs; EAL = English as an additional language.

and were given a standardized information sheet to read to participating children, including the aims of the study, confidentiality and the ability to withdraw at any time.

All items were administered to children online using a bespoke system designed to be easy to read and child-friendly with large font sizes. Recorded spoken accompaniment for all instructions, questionnaire items and response options was provided. The questionnaire items were presented to all students in the same order, with the bullying item preceding the school climate items.

Self-report measures were complemented with two sets of routinely recorded information. Child demographic information was obtained using the National Pupil Database which was linked to each participant. In addition, school level information was drawn from EduBase, a publically available database of school characteristics in England.

Instruments.

Bullying behavior. Participants reported on their own bullying behavior by indicating *never*, *sometimes*, or *always*, in response to the item "I bully others." This single item was included among a battery of measures (see Wolpert et al., 2011). Given only 2% of children responded *always* to this item, responses for *sometimes* and *always* were collapsed for all analyses (12%). As such, the measure indicates self-report of any bullying behavior, rather than the frequency of bullying behaviors.

To examine the validity of the single global bullying item two sets of analyses were conducted. First, children's self-reported bullying in the current study (12% of children report engaging in bullying at least sometimes) is comparable to the frequency of self-reported bullying reported in the literature (between 8% and 15% in late primary school/middle school pupils (e.g., Guerra et al., 2011; Nansel et al., 2001). Second, it is well documented that children who engage in bullying behaviors are also more likely to score highly on measures assessing externalizing behaviors (e.g., Cook, Williams, Guerra, Kim, & Sadek, 2010). As such, we examined whether this association holds for the current sample using the teacher-reported conduct problems subscale of the Strengths and Difficulties Questionnaire (Goodman, 1997), which was available for a nonrandom subset of the sample ($N = 2{,}197$) and the behavioral problems subscale of the Me and My School Questionnaire (Deighton et al., 2013), a validated self-reported measure of children's difficulties. Correlational analyses show a significant positive association between children's self-reported bullying using the single item and both teacher-reported ($r = .32$, $p < .001$) and self-reported ($r = .46$, $p < .001$) externalizing problems. For contrast, the correlation between the bullying item and children's teacher-rated ($r = .09$) and self-reported ($r = .015$) emotional problems were considerably lower.

Child demographic characteristics. Child characteristics included in analysis were gender, socioeconomic deprivation (FSM eligibility), ethnicity (White, Black, Asian, mixed, other/unclassified), special educational need (whether or not children were receiving special educational provision), language (whether or not English was an additional language for the child), and year group (Year 4 or Year 5).

School demographic characteristics. Routinely collected data at the school level included measures of school size (number of students), school gender (proportion of girls) and school deprivation (proportion of FSM eligible students). In addition, for each school we aggregated child-level data to estimate the school-level percentage of children from ethnic minority backgrounds (school ethnicity), percentage of children with a special educational need (school SEN) and percentage of children with English as an additional language (school EAL). Note that the latter therefore represent the demographic characteristics of the year group in question rather than the entire school. Table 1 presents descriptive information on the school characteristics.

School climate. All participants completed a self-report measure of school climate. School climate is a broad construct, and the current seven-item measure was derived specifically from measures of school supportiveness and school as a community (i.e., connectedness), including (a) staff support and care subscale of the School as Caring Community profile (Lickona & Davidson, 2003), (b) school-supportiveness subscale of the Sense of School Community Scale (Battistich & Hom, 1997), and (c) the My School scale of the Iowa Youth and Families Project Ratings Scale (Melby et al., 1993). For example, items included, "We can talk to teachers about problems" and, "At this school we care about each other." Participants responded by selecting one of three response options (*never*, *sometimes*, *always*). Scores ranged between 0 and 14, with higher scores indicating more positive perceptions of school climate. Cronbach's alpha (.75) demonstrates that the scale has adequate internal reliability. To ensure that these items were indeed assessing a single construct, we conducted principal component and factor analysis on the seven items. This analysis clearly indicated the existence of a single "school climate" factor with all items loading above 0.4 onto this single factor (Stevens, 1992). Average factor loading for individual items was 0.55, and ranged between 0.41 and 0.61.

Data analyses.

Missing data. The analyzed sample represents 95% of the possible 24,565 cases who were included in the study. Of the 1,350 cases that were excluded from analysis: 294 cases did not respond to the bullying item, 221 cases were excluded as they were missing child-level sociodemographic information (e.g., SEN, deprivation, language) in the National Pupil database and another 835 cases were excluded as there was no school level information available. A comparison of those who did ($n = 23{,}215$) and did not ($n = 294$) respond to the bullying item indicates that those who did not respond were significantly more likely to be of Asian (odds ratio [OR] = 1.65) or Black ($OR = 2.16$) ethnicity and be identified as having special educational needs ($OR = 1.19$). A comparison of those children who were missing the NPD data (demographics) indicates that a greater number of children with missing NPD data reported bullying (19%), compared to the 12% in the analyzed sample.

Statistical models. Following descriptive statistics, analyses were carried out in stages to estimate the amount of variation in bullying behavior accounted for by schools. To account for students nested within schools, all analysis was conducted using multilevel modeling with ML estimation. The models were constructed such that the child- and school- level predictors are modeled as fixed effects and we specify random effects at the school level as this is the cluster variable. To support interpretation of the interaction terms, we used grand mean centering to center the school- and child-level continuous variables in the models.

First, the baseline model (Model 1) was conducted that estimated overall school-level variance in bullying behavior. Second, child-level predictors were included in the model (Model 2). Third, school-level demographic (nonmalleable) predictors were added to the model (Model 3), so that the degree to which school composition variables are able to predict bullying behavior over and above child-level variables could be elucidated. Finally, school climate was included as a predictor of bullying behavior (Model 4). School climate was added on a separate step to determine if this malleable school-level factor predicts bullying behavior over and above nonmalleable, demographic school factors.

At each stage incremental model fit was estimated to assess if the additional predictors explained significantly more of the variation in bullying behaviors. For each model, we report the random effect parameter and the intraclass correlation (ICC) that represents the amount of variance in bullying accounted for by schools. By examining the ICC in consecutive models, the amount of variance previously attributed to schools, that is explained by the additional variables can be understood. Two further models were then run to examine child- by school-level interactions, Model 5A examines demographic interactions between children and schools, whereas Model 5B examines interactions between child-level variables and school climate.

Given the large sample size in the current study, the alpha rate for significance was set at $p < .01$ to minimize the likelihood of Type I error. All analyses were conducted in STATA version 12 (StataCorp, 2011).

Results

Results are presented in three parts. First, descriptive statistics are presented for key study variables. Second, multilevel logistic regression models with children nested within schools were used to explore child- and school-level predictors of children's self-reported bullying behavior. Finally, we explored the impact of interactions between child-level and school-level characteristics for predicting children's bullying behavior.

Descriptive statistics. Proportions of children responding *never* and *sometimes/always* to engaging in bullying behavior are presented in Table 2. There are several noteworthy features of Table 2. First, 17% of boys and 8% of girls report bullying others. Second, 18% of deprived children (i.e., those eligible for free school meals) report bullying others compared with 11% nondeprived children. Third, 21% of children with SEN reported bullying behavior in contrast with 9% children without SEN. In order to examine if these differences were statistically significant, multilevel logistic regression models were conducted and the results are reported in the following text.

Multilevel logistic regression models. The baseline model (Model 1; Table 3) indicates that 9.1% of the variance in bullying is accounted for by schools before any other child- and school-level variables were included. In Model 2 (see Table 3), including only child-level predictors of bullying behavior significantly improved the model, likelihood ratio test: D(9) = 974, $p < .001$, and explained an additional 1.4% of school-level variance in bullying behavior. Gender, deprivation, ethnicity and SEN were all found to be significant predictors of self-reported bullying behavior. Specifically, boys were more likely to report bullying behavior compared to girls, deprived children were more likely to report bullying compared to nondeprived children, and children from Black ethnic groups were more likely to report bullying compared to children from White backgrounds. Finally, children with a SEN classification were more likely to report bullying others compared to their peers without a SEN classification.

Table 2
Proportions of Children Responding "Never" and "Sometimes/Always" to the Bullying Behavior Questionnaire as a Function of Child-Level Demographic Characteristics

	Never (%)	Sometimes/ always (%)
Gender		
Male	82.98	17.02
Female	92.48	7.52
Deprivation (FSM)		
Not eligible	89.42	10.58
Eligible	81.94	18.06
Ethnicity		
White	88.28	11.72
Asian	86.92	13.08
Black	82.23	17.77
Mixed	86.44	13.56
Other/Not known	89.37	10.63
SEN		
No SEN	90.90	9.10
SEN	79.41	20.59
EAL		
No	87.94	12.06
Yes	86.54	13.46

Note. FSM = free school meals; SEN = special educational needs; EAL = English as an additional language.

In Model 3 (see Table 3), including demographic school-level predictors significantly improved the model, D(6) = 41.73, $p < .001$, explaining an additional 0.8% of school-level variance. The only significant school-level predictor in Model 3 was school deprivation, such that as the proportion of deprived children in the school increases, there was an increased likelihood of children reporting bullying behavior (over and above child-level FSM status). The pattern of significant child-level predictors remained unchanged from Model 2.

The inclusion of school climate in Model 4 (see Table 3), again significantly improved the model, D(1) = 110.41, $p < .001$, explaining an additional 2.5% of variance, over and above the variance explained by child- and school-level demographic factors. School climate was a significant independent predictor of bullying behavior, such that less positive perceptions of school climate was associated with greater self-reported bullying behavior. The pattern of child- and school-level characteristics remained unchanged with the addition of year group, which was now also a significant independent predictor of bullying behavior, children in Year 4 were likely to report greater bullying behavior compared to their older peers in Year 5.

Multilevel logistic regression models: Exploring Child × School interactions. Two additional models were also conducted to explore interactions between child- and school-level variables. In Model 5A (see Table 4), the incremental predictive power of child- by school-level demographic interactions were examined over and above the main effects. Specifically, this model explored whether children's individual demographic features in combination with school-level demographic characteristics predicted bullying behavior. Including demographic interaction significantly improved the model, D(8) = 30.93, $p < .001$, explaining

Table 3
Results of Multilevel Regression Models Testing Impact of Child- and School-Level Characteristics on Bullying Behavior

	Model 1: Baseline		Model 2: Child-level predictors		Model 3: School-level predictors		Model 4: School climate	
Parameter estimate	Estimate (*SE*)	*OR* (*SE*)	Estimate (*SE*)	*OR* (*SE*)	Estimate (*SE*)	*OR* (*SE*)	Estimate (*SE*)	*OR* (*SE*)
Child-level								
Intercept	−2.04 (.03)**	.13 (.004)	−1.61 (.10)**	.20 (.02)	−2.03 (.41)**	.13 (.05)	2.57 (.56)**	13.05 (7.33)
Gender (female)			−.86 (.05)**	.42 (.02)	−.86 (.05)**	.42 (.02)	−.86 (.04)**	.42 (.02)
Deprivation			−.42 (.05)**	1.52 (.07)	.35 (.05)**	1.43 (.07)	.37 (.05)**	1.44 (.07)
Ethnicity (Asian)			.17 (.10)	1.18 (.11)	.09 (.10)	1.10 (.11)	.12 (.10)	1.13 (.12)
Ethnicity (Black)			.42 (.09)**	1.53 (.14)	.33 (.09)**	1.39 (.13)	.32 (.09)**	1.38 (.13)
Ethnicity (Mixed)			.14 (.10)	1.15 (.12)	.10 (.10)	1.10 (.11)	.08 (.10)	1.09 (.11)
Ethnicity (Other)			−.05 (.16)	.95 (.16)	−.12 (.17)	.89 (.15)	−.09 (.16)	.91 (.15)
SEN			.75 (.04)**	2.12 (.09)	.72 (.04)**	2.06 (.09)	.72 (.04)**	2.06 (.09)
EAL			−.06 (.08)	.94 (.07)	−.17 (.08)	.84 (.07)	−.18 (.08)	.83 (.07)
Year group			−.16 (.06)	.85 (.05)	−.15 (.06)	.86 (.05)	−.28 (.06)**	.75 (.04)
School-level								
Size					−.03 (.02)	.97 (.02)	−.04 (.02)	.96 (.02)
Gender[a]					.06 (.08)	1.06 (.08)	.10 (.07)	1.10 (.08)
Deprivation[a]					.10 (.03)**	1.11 (.03)	.11 (.03)**	1.12 (.03)
Ethnicity[a]					−.01 (.03)	.99 (.03)	−.04 (.03)	.96 (.02)
SEN[a]					.03 (.02)	1.03 (.03)	.01 (.02)	1.01 (.0)
EAL[a]					.06 (.03)	1.06 (.03)	.07 (.03)	1.07 (.03)
School climate							−.39 (.04)**	.68 (.02)
Log-likelihood	−8546.51		−8059.34		−8038.48		−7983.27	
ICC	.091		.077		.069		.044	
Random effects	.57 (.03)		.52 (.03)		.49 (.03)		.39 (.03)	

Note. SEN = special educational needs; EAL = English as additional language.
[a] School composition characteristics are calibrated such that a unit represents 10%.
$^{*}p < .01$. $^{**}p < .001$.

an additional 0.2% of variance over and above Model 4 with child- and school-level main effects (see Table 4). The only interaction term that independently predicted bullying behavior was the child deprivation by school deprivation interaction, such that for nondeprived children the likelihood of reporting being a bully decreased with decreasing school-level deprivation. However, deprived children reported engaging in bullying irrespective of their school-level deprivation (see Figure 1). Including child by school climate interactions (Model 5B; Table 4) did not significantly improve the model, D(9) = 5.03, p = .083, and none of the interaction terms were significant (see Table 4). Finally, sensitivity analysis was conducted by excluding the special educational schools (n = 8) and single sex school (n = 1). Results remained unchanged.

Discussion

The current study examined a number of child demographic factors as well as malleable and nonmalleable school-level factors to better understand the predictors of bullying behavior across a large number of primary schools. Given much of bullying at this age takes place at school, it is important to understand both the independent influence of different school characteristics on the likelihood of bullying, as well as the combination of school- and child-level characteristics. Findings showed that both child- and school-level variables independently and in combination predicted children's bullying behavior. Specifically, boys, deprived children, those from Black ethnic groups, children with SEN and those from the younger year group were more likely to report bullying others. Over and above these child-level factors, increased school deprivation and poor school climate also predicted greater bullying behavior.

The current study also explored whether school-level factors moderate the association between child-level factors and bullying. Only the interaction between child deprivation and school deprivation was significant, such that deprived children were more likely to report bullying behavior regardless of the degree of deprivation of their school, while nondeprived children were more likely to report engaging in bullying behavior in schools with increased school-level deprivation. That is, being in a high-deprivation school is a risk factor for bullying only for children that come from nondeprived background. Untangling why this occurs requires further work, but it is possible that being a child from a nondeprived background in an otherwise deprived school sets up a peer group disparity or power imbalance that precipitates bullying behaviors. It would be interesting to examine other peer-related outcomes (e.g., friendship quality, victimization) to determine if this phenomenon is specific to bullying. This finding suggests that to understand the impact of deprivation on bullying behavior within a school it is crucial to take into account not only the degree of school deprivation but also the deprivation level of the individual child.

School climate also emerged as an important predictor of bullying behavior. The role of school climate for bullying behavior has been examined in a number of previous studies (e.g., Baker, 1998; Ma, 2002; Nansel et al., 2001) and has been established as being relevant for other related outcomes, such as mental health (Guerra et al., 2011). The current study's findings lend further support for this line of research. Current findings demonstrated

Table 4
Results of Multilevel Regression Models Testing Impact of Child- and School-Level Characteristics on Bullying Behaviour

Model 5a child × School interactions			Model 5b child × School climate interactions		
Parameter estimate	Estimate (*SE*)	*OR* (*SE*)	Parameter estimate	Estimate (*SE*)	*OR* (*SE*)
Child-level			Child-level		
Intercept	2.70 (1.00)		Intercept	4.78 (1.48)	
Gender (female)	−1.33 (.66)	.26 (.17)	Gender (female)	−1.79 (.68)**	.17 (.11)
Deprivation	.75 (.10)**	2.12 (.21)	Deprivation	.15 (.71)	1.17 (.82)
Ethnicity (Asian)	−.26 (.21)	.76 (.17)	Ethnicity (Asian)	−1.15 (1.42)	.32 (.45)
Ethnicity (Black)	.44 (.20)	1.56 (.31)	Ethnicity (Black)	.36 (1.37)	1.44 (1.97)
Ethnicity (Mixed)	.27 (.17)	1.31 (.23)	Ethnicity (Mixed)	−.94 (1.49)	.39 (.58)
Ethnicity (Other)	−.64 (.36)	.52 (.19)	Ethnicity (Other)	.14 (2.75)	1.16 (3.18)
SEN	.72 (.11)**	2.06 (.23)	SEN	.03 (.67)	1.03 (.69)
EAL	.02 (.15)	1.02 (1.5)	EAL	.52 (1.11)	1.68 (1.88)
Year Group	−.29 (.06)**	.75 (.04)	Year group	−1.12 (.86)	.34 (.28)
School-level			School-level		
Size	−.05 (.02)	.96 (.02)	Size	−.04 (.02)	.96 (.02)
Gender	.01 (.01)	1.06 (.09)	Gender	.01 (.01)	1.01 (.01)
Deprivation	.17 (.03)**	1.19 (.03)	Deprivation	.01 (.00)**	1.01 (.00)
Ethnicity	−.00 (.00)	.95 (.03)	Ethnicity	−.00 (.00)	1.00 (.00)
SEN	.00 (.00)	1.00 (.03)	SEN	.00 (.00)	1.00 (.00)
Language	.08 (.03)	1.08 (.03)	Language	.01 (.00)	1.01 (.00)
School climate	−.39 (.04)**	.68 (.02)	School climate	−.48 (.06)**	.62 (.04)
Interactions			Interactions		
Gender × School Gender[a]	.10 (.13)	1.10 (.15)	Gender × SC	.08 (.06)	1.08 (.06)
Deprivation × School Deprivation[a]	−.17 (.04)**	.85 (.03)	Deprivation × SC	.02 (.06)	1.02 (.06)
Ethnicity (Asian) × School Ethnicity[a]	.06 (.03)	1.06 (.04)	Ethnicity (Asian) × SC	.11 (.12)	1.12 (.14)
Ethnicity (Black) × School Ethnicity[a]	−.02 (.03)	.98 (.03)	Ethnicity (Black) × SC	−.00 (.12)	1.00 (.12)
Ethnicity (Mixed) × School Ethnicity[a]	−.05 (.04)	.95 (.03)	Ethnicity (Mixed) × SC	.09 (.13)	1.09 (.14)
Ethnicity (Other) × School Ethnicity[a]	.10 (.06)	1.10 (.06)	Ethnicity (Other) × SC	−.02 (.24)	.98 (.23)
SEN × School SEN[a]	−.00 (.03)	1.00 (.03)	SEN × SC	.06 (.06)	1.06 (.06)
EAL × School EAL[a]	−.05 (.03)	.95 (.03)	EAL × SC	−.06 (.10)	.94 (.09)
			Year Group × SC	.07 (.07)	1.07 (.08)
Log-likelihood	−7967.81		Log-likelihood	−7980.76	
ICC	.042		ICC	.043	
Random effects	.38 (.03)			.39 (.03)	

Note. SEN = special educational needs; EAL = English as additional language; SC = school climate; ICC = intraclass correlations.
[a] School composition characteristics are calibrated such that a unit represents 10%.
$^{*}p < .01$. $^{**}p < .001$.

that school climate is an important factor for understanding bullying in primary school over and above any child-level characteristics and nonmalleable school factors, highlighting the robust role played by school climate for peer relationships. Importantly, there were no significant interactions between school climate and child-level factors suggesting that the link with school climate is the same for all children in the school, regardless of their background of individual differences. It is important to note, however, that the current study is not able to determine the directionality of the findings between school climate and bullying, so it may be that children report poor school climate because of the degree of bullying in the school or, alternatively, poor school climate may be a factor driving bullying behaviors (Kasen et al., 2004). Further research exploring the impact of interventions to improve school climate on the incidence of bullying behavior for all students will be well placed examine the pattern of directionality between these constructs.

Given the power of the current study to detect significant effects both at the child- and school-level, it is notable that school size, school gender balance, ethnicity, SEN status and language all did not significantly independently predict bullying behavior in late primary school. This lends some support to other research with smaller samples of children that have also failed to find a significant association between these school composition factors and bullying (e.g., Saarento et al., 2013; Whitney & Smith, 1993).

Limitations

Although the current study has a number of strengths, notably a large sample of primary schools, a broad range of both child- and school-level indices, and an examination of the combined influence of child-level and school-level factors, there are several limitations to this work. First, the study's design was cross-sectional precluding an investigation of how child- and school-level factors may predict bullying behavior over time. As noted earlier, the directionality of influence between school climate and bullying behavior is unable to be determined from the current data.

Second, the measure of bullying comprised only a single self-report item, 'I bully others', with two response options (*sometimes* and *always*) collapsed. There was also no description of bullying provided to participants and a particular timeframe was not specified. While this is a clear limitation of the current study, using a

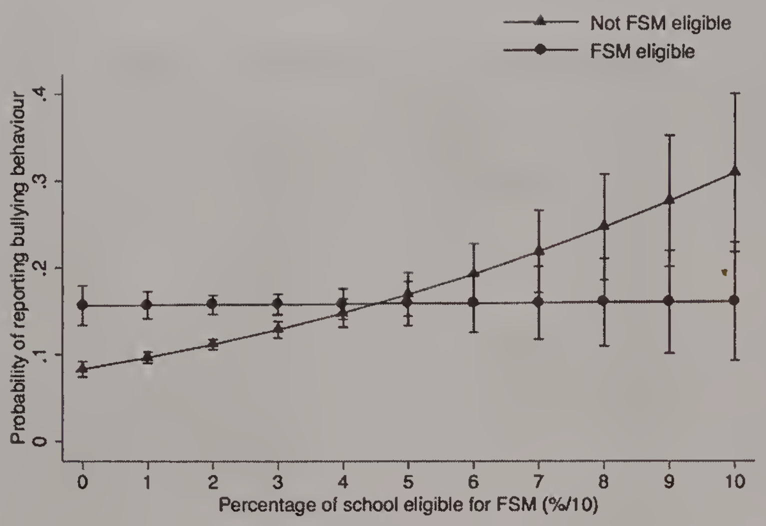


Figure 1. Predicted probabilities of reporting bullying behavior, showing interaction between child- and school-level deprivation. This figure represents fixed effects only.

single item to assess bullying behavior has been previously employed in the extant literature, especially in large scale national studies (e.g., Bradshaw et al., 2009; Nansel et al., 2001). Furthermore, the frequency of children's self-reported bullying in the current study (12% of children report engaging in bullying at least sometimes) is comparable to the frequency of self-reported bullying reported in the literature (between 8% and 15% in late primary school/middle school pupils (e.g., Guerra et al., 2011; Nansel et al., 2001). Nonetheless, a valid concern when using a single bullying item is that it may not have been sensitive to the nuances of different forms that bullying behavior may take, such as gossip, verbal bullying, and even cyber-bullying. Indeed, certain behaviors are more commonly perceived as bullying (i.e., physical bullying and name-calling, more male-typical bullying) compared to others (e.g., gossiping, exclusion) and may have resulted in girls underreporting bullying behaviors in the current study. Our results did show that boys reported more bullying than girls. However, this gender difference in the reporting of bullying behavior is a consistent feature of the bullying literature, even in those studies using more comprehensive self-reported bullying questionnaire measures (e.g., Pepler Jiang, Craig, & Connolly, 2008; Crick & Grotpeter, 1995), as well as single item bullying measures (e.g., Nansel et al., 2001), and peer-rated bullying nomination measures (e.g., Boulton & Smith, 1994). The fact that the current study also found this consistent gender difference using a highly abbreviated measure of bullying using only two categories of response (*never* vs. *sometimes/always*) lends support to the accuracy of both the single item and the response options. Nevertheless, further research using a large sample in conjunction with a more detailed measure of bullying behavior is clearly needed, and would allow for greater clarity on the association between child- and school-level characteristics and different forms of bullying behavior in children.

Finally, although the sample is large and representative of the wider population within participating schools, it does include greater number of children from deprived socioeconomic circumstances compared to all English primary schools. This makes it possible that the prevalence of bullying in the current study is an overestimate. However, given that socioeconomic status is controlled for in the analyses, we expect that the results pertaining to the child and school characteristics are robust.

Implications

The current findings have potential implications for the growing literature on how best to target school interventions to tackle bullying behavior (Smith, Ananiadou, & Cowie, 2003) and highlight the importance of targeting interventions, to both high-risk children and high-risk schools. In general, children from schools with a high proportion of children from more deprived backgrounds and with poorer school climate are at greatest risk of bullying behaviors. This suggests that promoting positive school climate through universal, whole-school approaches may be beneficial (Bosworth & Judkins, 2014). In addition, on the basis of the current study, it is clear that identifying the children that may be at risk of engaging in bullying behavior would be supported by considering not just the characteristics of the child (gender, deprivation, etc.) but also their relation to the wider school context (especially in terms of relative deprivation; e.g., Napoletano, Elgar, Saul, Dirks, & Craig, 2015). Future research using a similar approach might also investigate the interactions between social and cognitive individual characteristics such as cognitive ability, peer acceptance and school characteristics in predicting bullying and victim experiences. As such, this study may help to improve our ability to integrate whole-school and targeted antibullying programs, taking into account the school and child interactions that are associated with bullying, to allow more effective use of resources.

References

Baker, J. A. (1998). Are we missing the forest for the trees? Considering the social context of school violence. *Journal of School Psychology, 36,* 29–44. http://dx.doi.org/10.1016/S0022-4405(97)00048-4

Battistich, V., & Hom, A. (1997). The relationship between students' sense of their school as a community and their involvement in problem behaviors. *American Journal of Public Health, 87,* 1997–2001. http://dx.doi.org/10.2105/AJPH.87.12.1997

Bosworth, K., Espelage, D. L., & Simon, T. R. (1999). Factors associated with bullying behavior in middle school students. *The Journal of Early Adolescence, 19,* 341–362. http://dx.doi.org/10.1177/02724316990 19003003

Bosworth, K., & Judkins, M. (2014). Tapping into the power of school climate to prevent bullying: One application of schoolwide positive behavior interventions and supports. *Theory into Practice, 53,* 300–307. http://dx.doi.org/10.1080/00405841.2014.947224

Boulton, M. J., & Smith, P. K. (1994). Bully/victim problems in middle-school children: Stability, self-perceived competence, peer perceptions and peer acceptance. *British Journal of Developmental Psychology, 12,* 315–329. http://dx.doi.org/10.1111/j.2044-835X.1994.tb00637.x

Bowes, L., Arseneault, L., Maughan, B., Taylor, A., Caspi, A., & Moffitt, T. E. (2009). School, neighborhood, and family factors are associated with children's bullying involvement: A nationally representative longitudinal study. *Journal of the American Academy of Child and Adolescent Psychiatry, 48,* 545–553. http://dx.doi.org/10.1097/CHI.0b013e318 19cb017

Bradshaw, C. P., Sawyer, A. L., & O'Brennan, L. M. (2009). A social disorganization perspective on bullying-related attitudes and behaviors: The influence of school context. *American Journal of Community Psychology, 43*(3–4), 204–220. http://dx.doi.org/10.1007/s10464-009-9240-1

Bradshaw, C. P., Waasdorp, T. E., & Johnson, S. L. (2015). Overlapping verbal, relational, physical, and electronic forms of bullying in adolescence: Influence of school context. *Journal of Clinical Child and Adolescent Psychology, 44,* 494–508. http://dx.doi.org/10.1080/15374416.2014.893516

Bradshaw, C. P., Waasdorp, T. E., & Leaf, P. J. (2012). Effects of school-wide positive behavioral interventions and supports on child behavior problems. *Pediatrics, 130*(5), e1136–e1145. http://dx.doi.org/10.1542/peds.2012-0243

Brault, M.-C., Janosz, M., & Archambault, I. (2014). Effects of school composition and school climate on teacher expectations of students: A multilevel analysis. *Teaching and Teacher Education, 44,* 148–159. http://dx.doi.org/10.1016/j.tate.2014.08.008

Cook, C. R., Williams, K. R., Guerra, N. G., Kim, T. E., & Sadek, S. (2010). Predictors of bullying and victimization in childhood and adolescence: A meta-analytic investigation. *School Psychology Quarterly, 25,* 65–83. http://dx.doi.org/10.1037/a0020149

Crick, N. R., & Grotpeter, J. K. (1995). Relational aggression, gender, and social-psychological adjustment. *Child Development, 66,* 710–722. http://www.jstor.org/stable/1131945

Deighton, J., Tymms, P., Vostanis, P., Belsky, J., Fonagy, P., Brown, A., . . . Wolpert, M. (2013). The development of a school-based measure of child mental health. *Journal of Psychoeducational Assessment, 31,* 247–257. http://dx.doi.org/10.1177/0734282912465570

Department for Education. (2009). *Schools, pupils and their characteristics, January 2009.* London, UK: Stationery Office. Retrieved from http://webarchive.nationalarchives.gov.uk/20120504203418/http:/education.gov.uk/rsgateway/sc-schoolpupil.shtml

Espelage, D. L., & Swearer, S. M. (2003). Research on school bullying and victimization: What have we learned and where do we go from here? *School Psychology Review, 32,* 365–383.

Fink, E., Begeer, S., Hunt, C., & de Rosnay, M. (2014). False-belief understanding and social preference over the first 2 years of school: A longitudinal study. *Child Development, 85,* 2389–2403.

Fisher, H. L., Moffitt, T. E., Houts, R. M., Belsky, D. W., Arseneault, L., & Caspi, A. (2012). Bullying victimisation and risk of self-harm in early adolescence: Longitudinal cohort study. *British Medical Journal, 344,* e2683. http://dx.doi.org/10.1136/bmj.e2683

Glew, G. M., Fan, M. Y., Katon, W., Rivara, F. P., & Kernic, M. A. (2005). Bullying, psychosocial adjustment, and academic performance in elementary school. *Archives of Pediatrics & Adolescent Medicine, 159,* 1026–1031. http://dx.doi.org/10.1001/archpedi.159.11.1026

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 38,* 581–586. http://dx.doi.org/10.1111/j.1469-7610.1997.tb01545.x

Guerra, N. G., Williams, K. R., & Sadek, S. (2011). Understanding bullying and victimization during childhood and adolescence: A mixed methods study. *Child Development, 82,* 295–310. http://dx.doi.org/10.1111/j.1467-8624.2010.01556.x

Hale, D. R., Patalay, P., Fitzgerald-Yau, N., Hargreaves, D. S., Bond, L., Görzig, A., . . . Viner, R. M. (2014). School-level variation in health outcomes in adolescence: Analysis of three longitudinal studies in England. *Prevention Science, 15,* 600–610. http://dx.doi.org/10.1007/s11121-013-0414-6

Hawker, D. S. J., & Boulton, M. J. (2000). Twenty years' research on peer victimization and psychosocial maladjustment: A meta-analytic review of cross-sectional studies. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 41,* 441–455. http://dx.doi.org/10.1111/1469-7610.00629

Hobbs, G., & Vignoles, A. (2010). Is children's free school meal 'eligibility' a good proxy for family income? *British Educational Research Journal, 36,* 673–690. http://dx.doi.org/10.1080/01411920903083111

Kärnä, A., Voeten, M., Little, T. D., Alanen, E., Poskiparta, E., & Salmivalli, C. (2013). Effectiveness of the KiVa antibullying program: Grades 1–3 and 7–9. *Journal of Educational Psychology, 105,* 535–551. http://dx.doi.org/10.1037/a0030417

Kärnä, A., Voeten, M., Little, T. D., Poskiparta, E., Alanen, E., & Salmivalli, C. (2011). Going to scale: A nonrandomized nationwide trial of the KiVa antibullying program for Grades 1–9. *Journal of Consulting and Clinical Psychology, 79,* 796–805. http://dx.doi.org/10.1037/a0025740

Kasen, S., Berenson, K., Cohen, P., & Johnson, J. G. (2004). The effects of school climate on changes in aggressive and other behaviors related to bullying. In D. L. Espelage & S. M. Swearer (Eds.), *Bullying in American schools: A social-ecological perspective on prevention and intervention* (pp. 187–210). Mahwah, NJ: Lawrence Erlbaum.

Khoury-Kassabri, M., Benbenishty, R., Astor, R. A., & Zeira, A. (2004). The contributions of community, family, and school variables to student victimization. *American Journal of Community Psychology, 34*(3–4), 187–204. http://dx.doi.org/10.1007/s10464-004-7414-4

Kuperminc, G. P., Leadbeater, B. J., & Blatt, S. J. (2001). School social climate and individual differences in vulnerability to psychopathology among middle school students. *Journal of School Psychology, 39,* 141–159. http://dx.doi.org/10.1016/S0022-4405(01)00059-0

Lickona, T., & Davidson, M. L. (2003). *School as a caring community Profile-II (SCCP II).* Washington, DC: The Character Education Partnership.

Ma, X. (2002). Bullying in middle school: Individual and school characteristics of victims and offenders. *School Effectiveness and School Improvement, 13,* 63–89. http://dx.doi.org/10.1076/sesi.13.1.63.3438

Melby, J., Conger, R., Book, R., Rueter, M., Lucy, L., Repinski, D., & Ames, I. (1993). *The Iowa Family Interaction Rating Scales: Iowa Youth and Families Project. 4.* Ames, Iowa: Iowa State University.

Nansel, T. R., Overpeck, M., Pilla, R. S., Ruan, W. J., Simons-Morton, B., & Scheidt, P. (2001). Bullying behaviors among US youth: Prevalence and association with psychosocial adjustment. *Journal of the American Medical Association, 285,* 2094–2100. http://dx.doi.org/10.1001/jama.285.16.2094

Napoletano, A., Elgar, F. J., Saul, G., Dirks, M., & Craig, W. (2015). The view from the bottom: Relative deprivation and bullying victimization in Canadian adolescents. *Journal of Interpersonal Violence, 31,* 3443–3463. http://dx.doi.org/10.1177/0886260515585528

Olweus, D., & Limber, S. P. (2010). Bullying in school: Evaluation and dissemination of the Olweus Bullying Prevention Program. *American Journal of Orthopsychiatry, 80,* 124–134. http://dx.doi.org/10.1111/j.1939-0025.2010.01015.x

Pepler, D., Jiang, D., Craig, W., & Connolly, J. (2008). Developmental trajectories of bullying and associated factors. *Child Development, 79,* 325–338. http://dx.doi.org/10.1111/j.1467-8624.2007.01128.x

Saarento, S., Garandeau, C. F., & Salmivalli, C. (2015). Classroom- and school-level contributions to bullying and victimization: A review. *Journal of Community & Applied Social Psychology, 25,* 204–218. http://dx.doi.org/10.1002/casp.2207

Saarento, S., Kärnä, A., Hodges, E. V. E., & Salmivalli, C. (2013). Student-, classroom-, and school-level risk factors for victimization. *Journal of School Psychology, 51,* 421–434. http://dx.doi.org/10.1016/j.jsp.2013.02.002

Salmivalli, C. (2010). Bullying and the peer group: A review. *Aggression and Violent Behavior, 15,* 112–120. http://dx.doi.org/10.1016/j.avb.2009.08.007

Smith, P. (2010). Bullying in primary and secondary schools. In S. R. Jimerson, S. M. Swearer, & D. L. Esplage (Eds.), *Handbook of bullying in schools: An international perspective* (pp. 137–150). New York, NY: Routledge.

Smith, P. K., Ananiadou, K., & Cowie, H. (2003). Interventions to reduce school bullying. *Canadian Journal of Psychiatry, 48,* 591–599. http://dx.doi.org/10.1177/070674370304800905

StataCorp. (2011). *Stata statistical software: Release 12.* College Station, TX: Author.

Stevens, J. P. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Takizawa, R., Maughan, B., & Arseneault, L. (2014). Adult health outcomes of childhood bullying victimization: Evidence from a five-decade longitudinal British birth cohort. *The American Journal of Psychiatry, 171,* 777–784. http://dx.doi.org/10.1176/appi.ajp.2014.13101401

Vervoort, M. H., Scholte, R. H., & Overbeek, G. (2010). Bullying and victimization among adolescents: The role of ethnicity and ethnic composition of school class. *Journal of Youth and Adolescence, 39,* 1–11. http://dx.doi.org/10.1007/s10964-008-9355-y

Waasdorp, T. E., Pas, E. T., O'Brennan, L. M., & Bradshaw, C. P. (2011). A multilevel perspective on the climate of bullying: Discrepancies among students, school staff, and parents. *Journal of School Violence, 10,* 115–132. http://dx.doi.org/10.1080/15388220.2010.539164

Whitney, I., & Smith, P. K. (1993). A survey of the nature and extent of bullying in junior/middle and secondary schools. *Educational Research, 35,* 3–25. http://dx.doi.org/10.1080/0013188930350101

Wolke, D., Woods, S., Bloomfield, L., & Karstadt, L. (2000). The association between direct and relational bullying and behaviour problems among primary school children. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 41,* 989–1002. http://dx.doi.org/10.1111/1469-7610.00687

Wolke, D., Woods, S., Stanford, K., & Schulz, H. (2001). Bullying and victimization of primary school children in England and Germany: Prevalence and school factors. *British Journal of Psychology, 92*(Pt 4), 673–696. http://dx.doi.org/10.1348/000712601162419

Wolpert, M., Deighton, J., Patalay, P., Martin, A., Fitzgerald-Yau, N., Demir, E., . . . Frederikson, N. (2011). *Me and my school: Findings from the National Evaluation of Targeted Mental Health in Schools.* Retrieved from https://www.ucl.ac.uk/ebpu/docs/publication_files/tamhs_report

Received January 29, 2016
Revision received February 22, 2017
Accepted February 24, 2017 ■

Journal of Educational Psychology
2018, Vol. 110, No. 1, 27–45

0022-0663/18/$12.00 http://dx.doi.org/10.1037/edu0000205

Computer-Adaptive Testing: Implications for Students' Achievement, Motivation, Engagement, and Subjective Test Experience

Andrew J. Martin
University of New South Wales

Goran Lazendic
Australian Curriculum, Assessment, and Reporting Authority,
Sydney, Australia

The present study investigated the implications of computer-adaptive testing (operationalized by way of multistage adaptive testing; MAT) and "conventional" fixed order computer testing for various test-relevant outcomes in numeracy, including achievement, test-relevant motivation and engagement, and subjective test experience. It did so among N = 12,736 Australian elementary (years 3 and 5) and secondary (years 7 and 9) school students. Multilevel modeling assessed the extent to which Level 1 (student) test condition (fixed order vs. adaptive), gender, and year group factors and Level 2 (school) socioeducational advantage, location, structure, and size factors predicted students' test-relevant outcomes. In terms of statistically significant main effects, students in the computer-adaptive testing condition generated lower achievement error rates (i.e., higher measurement precision). Other statistically significant computer-adaptive test effects emerged as a function of year-level and gender, with positive effects of computer-adaptive testing being relatively greater for females and older students: these students achieved more highly (year 9 students), reported higher test-relevant motivation and engagement (year 9 students), and reported more positive subjective test experience (females and year 9 students). These findings (a) confirm that computer-adaptive testing yields greater achievement measurement precision, (b) suggest some positive test-relevant motivation and engagement effects from computer-adaptive testing, (c) counter claims that computer-adaptive testing reduces students' test-relevant motivation, engagement, and subjective experience, and (d) suggest positive computer-adaptive testing effects for older students at a developmental stage when they are typically less motivated and engaged.

Educational Impact and Implications Statement

With the growth in computer-based educational assessment, there are new opportunities to tailor testing to the characteristics and needs of students. Computer-adaptive testing is one such opportunity. Computer-adaptive testing presents items or sets of items (e.g., testlets) in a way to match different students' abilities. Based on a large-scale numeracy assessment exercise in Australia, our study showed that computer-adaptive testing assessed achievement accurately, had positive effects for key parts of students' test motivation and engagement, and benefited adolescent students who are at an age when they are typically less motivated and engaged. Computer-adaptive testing may therefore be a means to promote more positive experiences for students as they participate in computer-based and online educational testing.

Keywords: achievement, computer-adaptive testing, engagement, motivation, numeracy

Supplemental materials: http://dx.doi.org/10.1037/edu0000205.supp

With the growth in computer-based psycho-educational assessment, there are new opportunities to better tailor testing to the characteristics and needs of respondents. Computer-adaptive testing is one such opportunity gaining increasing attention and application. Whereas "conventional" fixed order tests administer items predominantly pitched to the "average" respondent, computer-adaptive testing administers items and/or item sets intended to be better matched to different respondents' abilities based on the individual's responses at the outset and through the test. It is timely to consider the implications of computer-adaptive testing for numerous test-relevant outcomes as

This article was published Online First April 20, 2017.

Andrew J. Martin, School of Education, University of New South Wales; Goran Lazendic, Australian Curriculum, Assessment, and Reporting Authority, Sydney, Australia.

We thank the Australian Curriculum, Assessment, and Reporting Authority (ACARA) for funding this research and Melissa McGilvray, Julie-Anne Justus, and Lifelong Achievement Group for assisting with the research.

It is appropriate to note that one of the measures (Motivation and Engagement Scale - Short) is from a published instrument attracting a small fee (approx. US$120 per 1000 respondents) that is put towards its ongoing development and administration and part of which is also donated to UNICEF. It is also appropriate to note that the second author is an employee of the agency funding this research.

Correspondence concerning this article should be addressed to Andrew J. Martin, School of Education, University of New South Wales, NSW 2052, Australia. E-mail: andrew.martin@unsw.edu.au

it is adopted more widely, including in national and statewide assessment programs. Based on a large-scale standardized numeracy assessment exercise using computer-adaptive and fixed order testing, the present study investigates three test-relevant outcomes: achievement, test-relevant motivation and engagement, and subjective test experience.

With regard to achievement, the study places some emphasis on the precision of achievement measurement. One of the contended yields of computer-adaptive testing is that better targeting of items to respondents should lead to less error within a test (Davey, 2011). Less intratest error is a desirable measurement property in itself (Lazendic & Adams, 2014; Wise, 2015) and is also considered desirable for the test-taker's experience of the test (Lifelong Achievement Group & Martin, 2015). With regard to test-relevant motivation and engagement, there has been mixed evidence when comparing computer-adaptive testing with more conventional fixed order tests (Colwell, 2013). This study therefore investigates the implications of computer-adaptive testing for test-relevant motivation and engagement because test-relevant motivation and engagement may affect respondents' willingness and capacity to demonstrate their knowledge and competence (Wise, 2015) and motivation and engagement are valued test-relevant outcomes in their own right (Colwell, 2013). Finally, with the predominant focus on computer-adaptive testing and its achievement effects, less attention has been given to respondents' subjective test experience (e.g., item comprehension, perceived test difficulty, perceived match to their ability etc.; Colwell, 2013; Wise, 2015). A positive subjective test experience is also a desirable end in itself (Lifelong Achievement Group & Martin, 2015) and also potentially linked to respondents' capacity to optimally demonstrate what they know and can do (Weiss, 2004). However, the limited research conducted into computer-adaptive testing is not clear on this. Taken together, focusing on these three test-relevant outcomes (achievement, motivation and engagement, subjective experience), we aim to build the very limited body of knowledge relevant to computer-adaptive testing and gain some clarity with regards to the mixed findings that have been evident to date.

Computer-Adaptive Testing

Nature of Computer-Adaptive Testing and Its Effects

Most psycho-educational assessment involves "conventional" tests that are paper-and-pencil or computer-based instruments in which items are presented to respondents in a predetermined, fixed, and linear order. They are typically designed using procedures of classical test theory (e.g., Cronbach, 1990; Gulliksen, 1950) such that items are selected using techniques that maximize the internal consistency (reliability) of the set of items that comprise the instrument (Weiss, 2004). Items are typically those that are appropriate for the "average" respondent, but are often too easy for respondents who are above average and too difficult for respondents who are below average on the factor being assessed (Wainer et al., 2000; Way et al., 2010; Weiss, 2004). Computer-adaptive testing aims to avoid asking questions that are much too difficult or much too easy for respondents (Davey, 2011). Instead, items or item sets are adaptively administered to respondents based on a respondent's answers at the outset and through the test. In so doing, the items better match the respondent's ability on the factor being measured.

There is research that supports computer-adaptive testing. Findings suggest that computer-adaptive testing tends not to adversely impact psychometric properties of test scores, enables greater measurement precision, can better accommodate students at more extreme ends such as low- and high-achievers, can build in test accommodations such as for students with disabilities, and can more appropriately reflect and assess the hierarchical nature of knowledge in a given subject area (e.g., Abedi, Leon, & Kao, 2007; Johnstone, Altman, & Thurlow, 2006; Minnema, Thurlow, Bielinski, & Scott, 2000; Stone & Davey, 2011; Thurlow, Lazarus, Albus, & Hodgson, 2010; Way et al., 2010). However, the research is not consistently positive. According to Colwell (2013), despite the improved measurement precision, respondents are still susceptible to the effects of test anxiety in computer-adaptive testing (see also Pitkin & Vispoel, 2001). Other research has identified a decrease in test-taker motivation and self-confidence in adaptive testing when compared to conventional testing (Frey, Hartig, & Moosbrugger, 2009; Häusler & Sommer, 2008). Given the increasing presence of computer-based testing in national, international, and other large-scale assessment exercises, the present study seeks to bring some clarity to the mixed findings derived to date.

Conceptual Underpinnings

Reviewing the research on computer-adaptive testing, it is evident that it is strongly underpinned by psychometric theory; connections to psycho-educational substantive theories are present, but relatively less developed. We also make the point that given the dearth of research in this area and its highly applied nature and implications, the study is not predominantly a "theory-testing" one and thus not one underpinned by hypotheses; rather this is a type of data-driven research that examines important research questions in education. Item response theory (IRT), in particular, has been particularly influential (de Ayala, 2009; Embretson & Reise, 2000; Stone & Davey, 2011). IRT is a means of placing items and students on the same scale, facilitating the direct matching of students to items that are most appropriate for them (Thompson & Weiss, 2011). In more technical terms, IRT is used to develop a set of overlapping "information function curves." Each curve indicates the range on the ability scale for which a corresponding item or item set will generate the most precise estimate of student ability. Matching items to students, based on IRT, is fundamental to computer-adaptive testing and its contended yields with regards to measurement precision and subjective test experience.

Computer-adaptive testing may also be consistent with current motivational theories and philosophies that focus on personalized learning, differentiation, and recognition of the hierarchical nature of knowledge and skills that students develop as they progress through school (Way et al., 2010). In line with these personalized and differentiated concepts, more is learned about the respondent by administering questions that challenge and scaffold test-takers, but do not overwhelm them (Davey, 2011). A related conceptual consideration is the notion of "fit" as relevant to the distinct and individual psycho-educational needs of young people. For example, stage-environment fit has attributed declining motivation and engagement to mismatches between adolescent competence and the extent to which high school accommodates and nurtures this competence and related needs (Eccles et al., 1993; Eccles & Roeser, 2009; Wang & Eccles, 2012). Computer-adaptive testing may offer better fit for different students at different levels of competence, allowing enhanced opportunities to stay engaged through the test and to demonstrate what they know and can do. Indeed, given the well-established decline in academic motivation and engagement into secondary school (in part attributed to the aforementioned poor fit; Booth & Gerard, 2014; Eccles et al., 1993; Gillen-O'Neel & Fuligni, 2013), computer-adaptive testing may play a useful role in testing students at this developmental stage.

Operationalization of Computer-Adaptive Testing

There are two typical computer-adaptive testing approaches. One is where the test difficulty is adjusted after each item. The other is where the difficulty of the test is adjusted after a student completes a set of items (Australian Curriculum, Assessment and Reporting Authority; ACARA 2014b; Lord, 1971a, 1971b, 1971c)—known as multistage adaptive testing (MAT). In MAT item sets are called testlets, with each testlet typically reflecting the composition of the complete test in terms of skills assessed and content. MAT is the approach investigated in the present study.

MAT has a number of advantages over item-level adaptive tests in the context of large-scale assessment programs because it provides better control over the administration and structure of the final tests and the capacity for students to preview and review items and to change their answers (see Hendrickson, 2007; Stone & Davey, 2011). In MAT, the content of all possible test pathways is determined in advance, enabling complete control over test curriculum coverage and coverage of knowledge, skills, and processes across the whole range of test difficulty. Such a control of the test content also enables MAT to eliminate the spurious dependencies in the test structure (where one item can provide cue for the subsequent item or items) by ensuring that items are placed in different test pathways. In addition, MAT is desirable because it can use a set of test items that depend on the same stimuli (such as a text passage in a reading test, e.g.), thus enabling more comprehensive and natural assessment of knowledge and skills in these test domains. Further, the number of items needed to implement MAT is typically smaller than that needed to implement item-level adaptive tests (ACARA, 2014b). However, these are contended operationalization yields and have not been the subject of much empirical attention. Whether they manifest in achievement, motivation and engagement, and subjective test experience benefits is an empirical question for this study.

Large-Scale Implementation of Computer-Adaptive Testing

Computer-adaptive testing now has many decades of technical research informing it, including research demonstrating comparability to paper-based or computer-administered conventional tests (Cudeck, 1985; Vispoel, Rocklin, & Wang, 1994) and the application of computer-adaptive testing to domain-specific assessment (Gibbons et al., 2008; Sands, Waters, & McBride, 1997; Thompson & Weiss, 2011).

The feasibility of large-scale computerized adaptive testing has been supported in numerous educational testing programs. For example, in the U.S, the adoption of the Common Core State Standards by numerous states has involved consortia funded to develop assessments aligned to the new standards that in part utilize computerized adaptive testing (Colwell, 2013; Smarter Balanced Assessment Consortia, 2015). Olson (2003) reports on other U.S states implementing statewide computer-adaptive testing aimed at more accurately measuring student achievement. Beyond school-based assessment exercises, licensing and certification assessment programs, such as U.S. Certified Public Accountants examinations (Breithaupt & Hare, 2007), also use adaptive testing.

In Australia (the site of this study), MAT has recently been piloted as the future national assessment program. Piloting has centered on assessment of numeracy for ACARA under its National Assessment Program—Literacy and Numeracy (NAPLAN). NAPLAN is the nation's standardized literacy and numeracy assessment administered annually to all students in years 3, 5, 7, and 9. In 2012, the Australian Government Department of Education funded ACARA to implement research to assist decision-making with regard to transitioning NAPLAN from paper-based form to adaptive computer (online) assessment (ACARA, 2014a). Over the past three years, ACARA and other investigators have conducted an ongoing research program to further develop and enhance the adaptive online NAPLAN test mode. Preliminary research has demonstrated the psychometric properties of the adaptive online test and its validity for use with a wide range of students and age groups (e.g., ACARA, 2013; Adams & Lazendic, 2013; Lazendic & Adams, 2014; Lowrie & Logan, 2013; Lifelong Achievement Group & Martin, 2015). The present study is the most recent in this program of research—investigating the achievement, motivation, engagement, and subjective experience implications of adaptive online testing.

Computer-Adaptive Testing and Implications for Achievement, Motivation, Engagement, and Subjective Experience

As computer-adaptive testing is adopted more widely, understanding its implications for numerous test-relevant outcomes is important for informing future educational assessment and policy (Colwell, 2013; Wise, 2015). The present study investigates three test-relevant outcomes: achievement, test-relevant motivation and engagement, and subjective test experience.

Implications for Test-Relevant Achievement

Adaptive tests aim to maximize the precision of test scores in relation to test length, or measurement efficiency (Davey, 2011). Thus, from a measurement perspective, adaptive tests can either improve score quality by being more precise than a conventional test of equal length or save time by being shorter than a conventional test of equal precision. Early evidence demonstrated improved measurement precision as reflected in less error within a test (e.g., Kingsbury & Weiss, 1983) without adversely impacting the psychometric characteristics of test scores (e.g., Mårdberg & Carlstedt, 1998; Moreno & Segall, 1997). Subsequent research has also suggested that computer-adaptive testing enables more precise measurement of achievement (Kingsbury & Hauser, 2004), including for low and high achieving students (ACARA, 2013; see also Davey, 2011; Stone & Davey, 2011; Lazendic & Adams, 2014).

Whereas achievement measurement precision is a potential yield of computer-adaptive testing, computer-adaptive testing is not aimed at raising scores. Thus, although there should be greater measurement precision (reflected by less intratest error variability) in computer-adaptive testing when compared with a fixed order computer test, there should not be differences in achievement levels between computer-adaptive testing and fixed order computer testing. We therefore compare achievement level and achievement measurement precision for students taking computer-adaptive testing and students taking a fixed order computer test. We should make the point that of all the effects investigated in this study, we deem significant achievement precision effects and nonsignificant achievement level effects as the more likely effects. To the extent that this is the case, the more novel and exploratory aspect of the investigation is the role of computer-adaptive testing in test-relevant motivation, engagement, and subjective test experience.

Implications for Test-Relevant Motivation and Engagement

A small body of work has explored the potentially motivating and engaging properties of computer-adaptive testing. However, as noted above, the limited research conducted into adaptive testing and test-relevant motivation and engagement is not consistent. Colwell (2013) points out that respondents are still susceptible to the effects of test anxiety in adaptive testing (see also Pitkin & Vispoel, 2001), whereas other researchers identify declines in test-taker motivation in adaptive testing (Frey et al., 2009; Häusler & Sommer, 2008). Also, if motivation is improved in adaptive testing, presumably this should positively impact achievement scores and yet this is not a widely noted outcome in high quality adaptive testing. In sum, the claims for and against the motivation and engagement implications of computer-adaptive testing have yet to receive a level of empirical attention to warrant sufficiently robust conclusions. Addressing these claims is another major purpose of the present investigation.

Motivation is here defined as individuals' energy, inclination, interest, and drive to learn, work effectively and achieve to potential—and engagement as the behaviors aligned with or following from this energy, inclination, interest, and drive (Martin, 2007, 2009). The present investigation utilizes a recently proposed motivation model—the Motivation and Engagement Wheel—as a basis for examining the motivation- and engagement-related implications of computer-adaptive testing. The Wheel comprises positive and negative motivation and engagement factors. Positive motivation and engagement is composed of self-efficacy, valuing, mastery orientation, planning and monitoring, task management, and persistence. Negative motivation and engagement comprises anxiety, failure avoidance, uncertain (low) control, self-handicapping and disengagement (for further details, see Materials for construct operationalization). Given there is no "grand theory" of motivation and engagement (Pintrich, 2003; see also Reeve, 2016), these 11 motivation and engagement factors have been identified based on their integration of seminal conceptualizing such as self-efficacy, attribution and control, valuing, achievement goal orientation, need achievement, self-worth, self-determination, and self-regulation theories (see Liem & Martin, 2012 for a review of these theories as relevant to the Wheel). Alongside the Wheel is its accompanying instrumentation, the Motivation and Engagement Scale (MES; Martin, 2014), a well-established measure (first validated by Martin, 2001; see also a review of its history and psychometrics by Liem & Martin, 2012) that assesses each of the 11 motivation and engagement factors in the Wheel. The MES has been applied in different domains, including in mathematics, science, and English subjects (Green, Martin, & Marsh, 2007; Martin, Anderson, Bobis, Way, & Vellar, 2012). There is thus empirical support to use the MES in a domain-specific way, that, in the case of the present investigation involved adaptation to assess test-relevant motivation and engagement (see Method below).

Implications for Subjective Test Experience

Alongside potential implications for achievement, motivation, and engagement, better targeted items may also lead to a more positive subjective experience of the test itself. Preliminary research suggests that adaptive testing caters more fully to students' assessment and learning needs (ACARA, 2013) and has potential to enable students to complete the test with a greater sense of satisfaction (Lowrie & Logan, 2013). It is also possible that when items do not optimally match students' ability, other factors are introduced that compromise measurement precision (Weiss, 2004), including confusion and discomfort. Indeed, research has demonstrated that some of the adverse effects of high-stakes testing on students include illness and heightened levels of stress and tension (Mulvenon, Stegman, & Ritter, 2005; Triplett, Barksdale, & Leftwich, 2003); thus, better matching of items to students may alleviate these negative subjective experiences.

In this study we investigate numerous indicators of subjective test experience, including perceived ease of test items, the extent to which students feel items match their ability, students' interest in test items, item comprehension, and students' perceived performance. Given the claims about the benefits of better matching items to respondents (e.g., Stone & Davey, 2011), we might expect students in the computer-adaptive test condition to report more favorably on their subjective test experience than students in the fixed order computer test condition. However, very little research has examined this issue. Although there may be positive effects of computer-adaptive testing for subjective test experience, it remains an empirical question that is a further major focus of the present study.

Covariates Important to Consider

This study's test-relevant outcomes (achievement, motivation, engagement, and subjective experience) may also be correlated with various student-level and school-level background factors. We therefore include such factors as covariates in order to partial out their variance and better understand the unique effects of computer-adaptive testing on the test-relevant outcomes of interest. Inclusion of these covariates also allows us to assess their effects on test-relevant outcomes as a research question in its own right. Student-level background factors are gender and year-level.[1] Four school-level background factors are of particular interest in the present study: school socioeducational advantage, school size, school location, and school structure. Because these are not central to the research questions in our study, a detailed conceptual and

[1] As a nationally based assessment exercise administered by a central authority and subject to restrictions on access to personal information, limited student-level personal data were available to the researchers for analysis.

empirical rationale for inclusion of these covariates is provided in Supplementary Materials.

The Present Study

This investigation explores the implications of computer-adaptive testing (in the form of multistage adaptive testing [MAT]) and "conventional" fixed order computer testing for various test-relevant outcomes in the numeracy domain. The study is conducted within a multilevel context and assesses the extent to which (a) test condition (fixed order vs. adaptive), gender, and year group (Level 1; student-level) factors and (b) school socioeducational advantage, school location, school structure, and school size (Level 2; school-level) factors predict students' achievement, test-relevant motivation and engagement, and subjective test experience in a national standardized online numeracy test. Figure 1 shows the proposed analytic model. The study is centrally concerned with understanding the effect of computer-adaptive testing (by way of MAT) on students' test-related outcomes after accounting for shared variance among various student- and school-level predictors (covariates). Findings, we suggest, hold implications for test administrators, test validity, and for promoting more positive experiences for students as they participate in large-scale computer-based and online testing.

Method

Participants

For the present study, the Australian Curriculum, Assessment and Reporting Authority (ACARA) invited the participation of all Australian states, territories, and school systems organized in eight Test Administration Authorities (TAA). Each TAA received a request to provide a designated number of schools to ensure proportional representation relative to the number of schools in that TAA. They were asked to ensure that sampled schools covered the range of school types, geographical location, and school achievement. All TAAs initially sought expressions of interest from schools under their jurisdictions, but to ensure the representativeness of the sample and diversity of students, some TAAs worked with systems and sectors to nominate and mandate participation of some school types in the study. Consequently, schools that TAAs nominated for the study were similar in characteristics to the other schools that volunteered participation in the study. The proportion of nominated schools varied with TAA and school sector and therefore there was no systematic bias introduced in the sample owing to the process of nomination and selection of schools in the sample.

The final sample comprised 231 schools from all Australian states and territories, from all sectors (government, Catholic, and independent), and from urban (where the vast majority of schools are located), semiurban (regional), remote, and very remote locations. Appendix A shows the composition of sampled schools relative to all Australian schools. It will be noted that to gain adequate numbers in some areas, some small TAAs (e.g., the Australian Capital Territory and Northern Territory) were slightly overrepresented in the sample and some highly diverse TAAs in large geographic areas (e.g., New South Wales) were also slightly overrepresented to ensure coverage of different students across a large geographic area. Further details on selection of classes within schools is provided in Supplementary Materials.

Participants were 12,736 elementary (years 3 and 5) and secondary (years 7 and 9) school students. They were from year 3 (N = 3,557; 28%), year 5 (N = 3,797; 30%), year 7 (N = 2,911; 23%), and year 9 (N = 2,471; 19%). Approximately 50% were females (N = 6,392), with N = 6,344 males. Two-thirds of schools (N = 154; 67%) were urban, 28% (N = 64) were semiurban, 3% (N = 8) were remote, and 2% (N = 5) were very remote. Just under 60% (N = 136; 59%) were elementary schools, 23% (N = 53) were secondary schools, and 18% (N = 42) were combined elementary and secondary schools. The mean ICSEA (school socioeducational advantage; national average = 1000) score for schools was 1016 (SD = 83). The average school size for elementary schools was N = 383 (SD = 246) students, for secondary schools it was N = 890 (SD = 393) students, and for combined schools it was N = 885 (SD = 571) students. The fixed test condition comprised N = 7,152 students (56% of the sample) and the adaptive test condition comprised N = 5,584 students (44% of the sample). Table 1 presents sample details for each of the fixed and adaptive test conditions as a function of key subgroups. Table 1 also provides descriptives for the fixed and adaptive test conditions as relevant to the central factors in the study (achievement, motivation and engagement, subjective experience).

Procedure

Test administration. Within each school, students were allocated to either the fixed or the adaptive test condition. To achieve the allocation of students across conditions, two sets of test login codes were created, one for each test condition. All test sessions were supervised by external proctors engaged by ACARA (see Supplementary Materials for further detail on Proctors' role in schools). Proctors distributed the login codes to each student at the start of the test session using a "round-robin" approach. In this approach, a first student in the class received the login code for the adaptive test condition, the second student in a class received the login code for the fixed condition, the third for the adaptive condition, the fourth for the fixed condition, and so on. Students were assigned their login code in the order in which they entered the classroom, in the order of their seating in the classroom, or in alphabetical order. The proctors were then instructed to change the order of login code distribution in a subsequent class or school that they visited. In this way, students were allocated to test conditions, with equal proportions, without being aware that they were taking tests under two different conditions. As the study was done in conjunction with the trialing of new items for future online tests and in order to achieve the required sample size for the trialing of new items, the proportion of students who received tests in the fixed condition was increased in some schools—as described in Supplementary Materials.

Implementation of MAT. The adaptive form of NAPLAN for this study implemented a MAT design. In MAT, the test difficulty is adjusted after a student provides responses to a set of items; thus a student progresses through a series of stages containing item sets (testlets). Students passed through three stages containing testlets of varying difficulties. The test pathways were determined by the two branching points. All students started with the same set of items (testlet A). Based on their performance in the

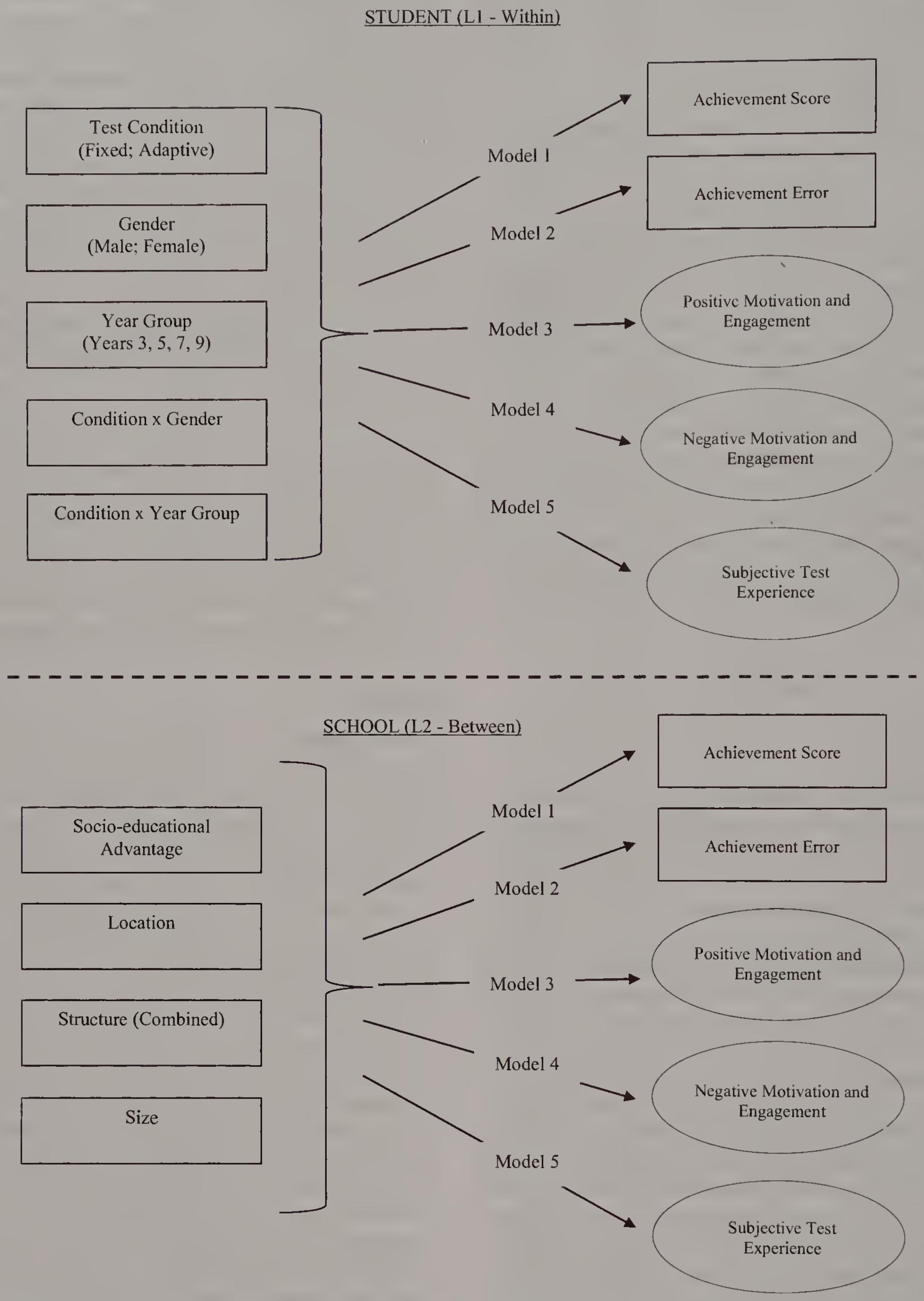

STUDENT (L1 - Within)
Test Condition (Fixed; Adaptive)
Gender (Male; Female)
Year Group (Years 3, 5, 7, 9)
Condition x Gender
Condition x Year Group
Model 1
Model 2
Model 3
Model 4
Model 5
Achievement Score
Achievement Error
Positive Motivation and Engagement
Negative Motivation and Engagement
Subjective Test Experience
SCHOOL (L2 - Between)
Socio-educational Advantage
Location
Structure (Combined)
Size
Model 1
Model 2
Model 3
Model 4
Model 5
Achievement Score
Achievement Error
Positive Motivation and Engagement
Negative Motivation and Engagement
Subjective Test Experience

Figure 1. Proposed analytic model exploring effects of computer-adaptive testing. Although not included in the figure, for each of the two levels, there is a predictive path between each independent variable and each dependent variable. Thus, for example, Model 1 has five Level 1 predictors of Achievement Score and four Level 2 predictors of Achievement Score.

testlet A, students were branched to a second testlet. The second testlet may be easier (B) or more difficult (D) than testlet A. At the end of the second testlet, students were directed to a third testlet, this time depending on achievement in both first and second testlets. The final testlets were of varying difficulty: hard (F) containing the most challenging set of items, medium (E) containing the mainstream set of items, and easy (C) containing items designed to elicit information about performance of students that were not progressing as expected. Each testlet contained approximately one third of the total number of items in the overall test

Table 1

Fixed and Adaptive Conditions: Sample and Factor Descriptives

			Achievement score		Achievement error		Positive motivation and engagement		Negative motivation and engagement		Subjective test experience	
Descriptive	Fixed condition *N*	Adaptive condition *N*	Fixed condition *M* (*SD*)	Adaptive condition *M* (*SD*)	Fixed condition *M* (*SD*)	Adaptive condition *M* (*SD*)	Fixed condition *M* (*SD*)	Adaptive condition *M* (*SD*)	Fixed condition *M* (*SD*)	Adaptive condition *M* (*SD*)	Fixed condition *M* (*SD*)	Adaptive condition *M* (*SD*)
Year-level												
Year 3	2048	1509	.01 (.97)	−.01 (1.04)	.44 (.11)	.40 (.05)	5.37 (1.11)	5.33 (1.11)	2.83 (1.27)	2.84 (1.23)	NA	NA
Year 5	2070	1727	−.01 (.99)	.01 (1.01)	.42 (.12)	.37 (.04)	5.34 (.98)	5.31 (.95)	2.70 (1.18)	2.74 (1.10)	NA	NA
Year 7	1624	1287	−.03 (.93)	.04 (1.08)	.46 (.08)	.38 (.06)	4.80 (1.11)	4.84 (1.02)	2.87 (1.10)	2.82 (1.03)	4.50 (.97)	4.59 (.86)
Year 9	1410	1061	−.07 (.98)	.36 (.05)	.41 (.11)	.36 (.05)	4.22 (1.17)	4.29 (1.12)	3.02 (1.07)	2.82 (1.03)	4.14 (.97)	4.35 (.84)
Gender												
Male	3588	2756	.02 (1.01)	.04 (1.11)	.44 (.12)	.43 (.10)	5.03 (1.18)	5.04 (1.16)	2.74 (1.19)	2.74 (1.13)	4.45 (.96)	4.56 (.88)
Female	3564	2828	−.07 (.93)	.01 (.96)	.38 (.05)	.37 (.05)	4.97 (1.16)	4.99 (1.07)	2.94 (1.15)	2.89 (1.08)	4.23 (.99)	4.42 (.84)
School												
Elementary	3870	2986	−.04 (1.25)	.02 (1.30)	.43 (.11)	.38 (.05)	5.28 (1.06)	5.28 (1.01)	2.79 (1.23)	2.80 (1.15)	4.52 (1.04)	4.70 (.79)
Secondary	1654	1442	−.29 (1.18)	−.17 (1.31)	.42 (.09)	.37 (.05)	4.42 (1.14)	4.52 (1.07)	2.96 (1.08)	2.85 (1.02)	4.28 (.94)	4.43 (.85)
Combined	1628	1156	−.03 (1.32)	.01 (1.42)	.44 (.11)	.38 (.05)	4.92 (1.18)	4.93 (1.14)	2.84 (1.16)	2.82 (1.11)	4.34 (.99)	4.48 (.88)
Location												
Urban	4939	3786	.06 (.97)	.12 (1.06)	.44 (.11)	.38 (.06)	5.05 (1.15)	5.04 (1.12)	2.78 (1.15)	2.81 (1.15)	4.41 (.97)	4.53 (.86)
Semi-urban	1993	1641	−.17 (.94)	−.13 (.94)	.43 (.10)	.37 (.04)	4.91 (1.20)	4.96 (1.12)	2.96 (1.21)	2.82 (1.11)	4.19 (1.00)	4.39 (.85)
Remote	127	119	−.55 (.89)	−.44 (1.04)	.43 (.08)	.37 (.04)	4.67 (1.38)	4.93 (1.06)	3.10 (1.36)	3.09 (1.07)	4.05 (.98)	4.45 (.99)
Very remote	93	38	−.52 (.99)	−.65 (1.42)	.44 (.08)	.39 (.09)	5.09 (1.03)	5.24 (.80)	3.15 (1.13)	2.97 (.90)	4.26 (1.05)	4.48 (1.36)
Total	7152	5584	−.02 (.97)	.03 (1.04)	.43 (.11)	.38 (.05)	5.00 (1.16)	5.01 (1.10)	2.84 (1.18)	2.82 (1.11)	4.34 (.98)	4.48 (.85)

Note. NA = Year 3 and Year 5 were not administered these items to reduce the time burden on these younger students.

and was additionally representative of the whole test in terms of skills and content coverage. Figure 2 shows the test paths available to students in this MAT design, including a test pathway A-C-B that is designed to cater for students who could not successfully engage with the initial testlet A. Students who cannot answer more than two or three items correctly in testlet A are sent directly to the easiest testlet C to increase their chances of engagement with the rest of the test. Upon completion of testlet C these students are sent to testlet B to give them an opportunity to show the full range of their ability. Supplementary Materials provides further detail on the MAT algorithms used in the computer-adaptive testing.

In the fixed condition, students were randomly administered one of four linear, not adaptive, tests. The four linear tests corresponded to the four test pathways available to students in the adaptive test condition (A-D-F, A-D-E, A-B-E, A-B-C). The four linear tests were randomly allocated to students in the fixed test condition using the relevant login codes and "round robin" method for allocation of login codes as described above. The review of items and changes to the item response were enabled within each testlet. Before moving to the next testlet, students received the message informing them that they should check their response before proceeding and that once they proceed to the next part of the test they will not be able to return to items in the previous test stage. The same functionality and instructions were provided to students in both adaptive and fixed test conditions thus further ensuring that interaction with test items and content was the same across the two experimental conditions.

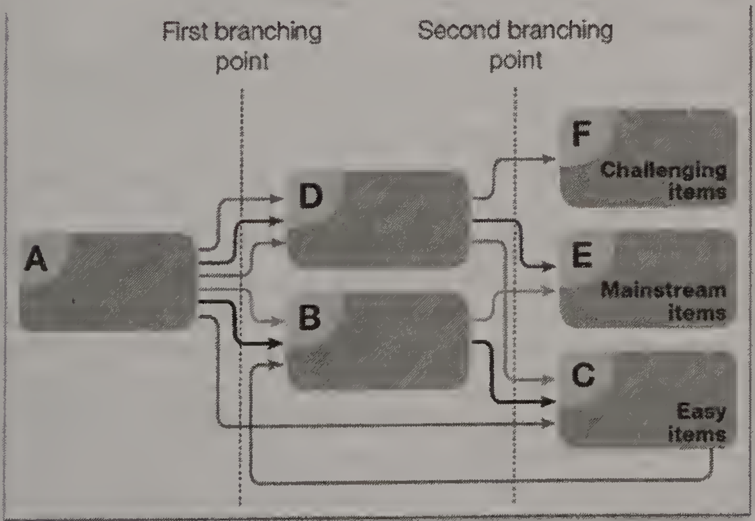


Figure 2. Computer-adaptive test (by way of multistage adaptive testing [MAT]) design: Testlets and six test pathways available (ACARA, 2014c).

Materials

NAPLAN testlets and tests: An overview. Current NAPLAN tests reflect the construct of numeracy as described in Statements of Learning for Mathematics (Ministerial Council on Education, Employment, Training, & Youth Affairs, 2006). The statements define numeracy as mathematical knowledge, skills and understanding against five broadly defined and interrelated aspects of mathematics curricula that are considered essential and common across Australia. These five subdomains are as follows: (a) number; (b) algebra, function, and pattern; (c) measurement, chance, and data; (d) space; and (e) working mathematically (comprising mathematical reasoning, mathematical problem solving, and communicating mathematical solutions). Each item in current numeracy tests is explicitly mapped against one of the first four domains and "working mathematically" is applied as a general concept to construct items and tests that are assessing numeracy and not just mathematics as a curriculum domain. The numeracy tests contain two types of items: multiple-choice and constructed response.

The development of NAPLAN tests has been guided by the Rasch model (Rasch, 1960/1980). The Rasch model has the assumption that discrimination parameters (or factor loadings) are held constant across items. When Rasch model analyses show that an item does not fit a unidimensional measurement model, this item is not included in the final test. Consequently, NAPLAN tests generate a single Rasch model estimate of student numeracy. Equally important, the comparability of ability measures in NAPLAN is maintained across tests for different year-levels as NAPLAN tests are vertically linked through a set of common link items. Further details on the numeracy items and testlets for this study are presented in Supplementary Materials.

The fixed test also included the newly developed parallel forms of all testlets used in the adaptive test condition. The parallel testlets have been created to mirror the range of item difficulty and content and knowledge and skills coverage of the testlets used in the adaptive test condition. The four linear forms were then created by combining the testlets used in the adaptive condition and the new parallel testlets. For example, in the adaptive condition, testlet A was used in two linear test forms and a parallel version of testlet A was used in the remaining two linear forms. Similarly, one of each version of testlet B, D, and E was used and only a parallel version of testlet C and F have been used. The functioning of the new parallel testlets have been examined in the analyses phase of the research and Rasch model analyses show that these new items have a satisfactory fit to the model (Lazendic & Adams, 2014) and thus there were no impediments in using these items and testlets in the analyses of students' performance in the fixed condition.

Regardless of test condition all testlets had 12 items in year 3, 14 items in year 5, and 16 items in years 7 and 9. Consequently the total number of item in a test was 36, 42, and 48, respectively. Some of the year 7 and 9 items required the use of a calculator which was provided as an onscreen solution only for those items.

Outcome measures.

NAPLAN test (achievement and error). As noted above, NAPLAN *achievement* is a score based on measurement scales constructed using the Rasch model (Rasch, 1960/1980). Each student also received an *achievement error (or, precision) score*, a standard error of WLE ability estimate, which was the variability in correct responding from item-to-item across the total test items for each student. Thc advantage of the Rasch model is that it enables algebraic separation of the person ability and item difficulty parameter estimations. That is, the person ability can be eliminated during the process of statistical estimation of the item difficulty and vice versa. The final outcome of both measurement models is a single, interval, scale on which item difficulty and person ability parameters can be located. In the context of MAT, this means that students who took tests consisting of different items can have their abilities estimated using the same scale as long as there are some shared items between different tests. The same principle applies for tests used in the fixed and adaptive condition as there was a significant number of testlets shared between tests in the two conditions.

The Rasch model analyses showed that almost all items have fit statistics (infit mean square) inside of the 0.7–1.3 range that indicates productive measures and items that fit the unidimensional measurement model (for an overview of fit statistic recommendations, see Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008). The infit mean square chi-square statistic has each observation weighted by its contribution to the model variance. The infit mean square has the distribution of an expected value of 1 and thus items with fit statistics considerably different from 1 can be considered to be showing more variation between the model and the observed scores than predicted by the unidimensional Rasch model. Items with infit mean square significantly higher than 1 (viz.,1.3) are considered to be less discriminating relative to the model and items with fit statistics lower then 1 (viz., 0.7) are considered to be more discriminating relative to the unidimensional model fitted to the test response data. Across all year-levels there were no items that had infit mean square lower than 0.7. In year 3, there were only three items that had infit mean square larger than 1.3, in year 5 there was one such item, in year 7 six items, and in year 9 there were 11 items with infit mean square larger than 1.3. The proportion of items that did not fit the model for years 3, 5, and 7 (relative to the full item set across all testlets) is very low at 4%, 1%, and 7% respectively. The proportion of misfitting items in year 9 is 11%, but these misfitting items are spread across all test content areas which means that there are no indications that misfit of individual items is violating the unidimensionality, and thus the construct validity of the whole test. Moreover, and importantly for the MAT design, misfitting items are not concentrated in a single testlet and therefore the progression of students through adaptive tests has not be materially affected by the misfitting items. The same findings were observed for items and tests in other years, including year 7, and thus we deemed it defensible to proceed with the full item set for all years.

Once the process of item difficulty calibration and inspections was completed, the obtained item difficulty parameters were used to estimate the students' ability using the weighted likelihood estimation (WLE) procedure. The scaling was done separately for each year-level and to further ensure comparability across year-levels for analyses in the study. Because there were differences between year-levels in achievement scores (year 3 M = 0.02, SD = 1.26; year 5 M = 0.15, SD = 1.29; year 7 M = −0.62, SD = 1.13; year 9 M = 0.12, SD = 1.31), students' NAPLAN achievement score was standardized within each year-level such that each of years 3, 5, 7, and 9 had a mean of 0 and SD of 1. Means and SDs for adaptive and fixed conditions are shown in Table 1. Reliability of the achievement scores in this study was high. The WLE person separation indices were .88, .90, .91, and .91 for years 3, 5, 7 and 9, respectively. Adams (2005) has shown that, in the context of the Rasch model, the person separation index is functionally equivalent to the Kuder–Richardson formula 20 (KR-20) reliability coefficient. In variance components analysis (null model), the ICC (proportion of school-level variance) for achievement score was .23 and for achievement error was .07.

Test-relevant motivation and engagement. Upon completion of the NAPLAN test items, students responded to a set of items assessing their test-relevant motivation and engagement—that is, their perceived motivation and engagement while doing the test. Motivation and engagement were assessed using the Short Motivation and Engagement Scale (Short MES; Martin, 2014), adapted for the online NAPLAN test. The Short MES has been successfully adapted for research in other domain-specific research (e.g., the arts; Martin et al., 2013) and so was deemed appropriate for adaptation here. All Short MES items were rated on a 1 (*Strongly Disagree*) to 7 (*Strongly Agree*) scale.

Positive (or, adaptive) test-relevant motivation and engagement comprises six items, *self-efficacy* (“I did well in this test”), *mastery orientation* (“In this test I was very focused on understanding the questions and tasks”), *valuing* (“This test was important”), *persistence* (“I persisted in this test even when it was challenging or difficult”), *planning* (“In this test, I planned my answers and monitored my progress”), and *task management* (“In this test I made good use of my time”). As described below, using M*plus*, a latent factor was formed using these six indicators (M = 5.01, SD = 1.14, Cronbach’s alpha = .78, mean factor loading = .64). In variance components analysis (null model), the ICC (proportion of school-level variance) for positive motivation and engagement was .07. Means and *SD*s for adaptive and fixed conditions are shown in Table 1.

Negative (or, maladaptive) test-relevant motivation and engagement comprises five items, *anxiety* (“I was anxious in this test”), *failure avoidance* (sometimes referred to as performance avoidance; “In this test I did not want to get a bad mark”), *uncertain (low) control* (“I don’t think I had much control over how well I did in this test”), *self-handicapping* (“During this test I wasted time and was easily distracted”) and *disengagement* (“I often felt like giving up in this test”). A latent factor was formed using four of these indicators (M = 2.83, SD = 1.15, Cronbach’s alpha = .60, mean factor loading = .51; failure avoidance was dropped as its inclusion yielded reliability < .60). In variance components analysis (null model), the ICC (proportion of school-level variance) for negative motivation and engagement was .04. Means and *SD*s for adaptive and fixed conditions are shown in Table 1.

Given its marginal reliability, we also conducted a multilevel (students at Level 1; schools at Level 2) confirmatory factor analysis for the negative motivation and engagement factor (see Analyses, below, for description of fit indices used here). The fit was very good, $\chi^2(4) = 206$, CFI = .96, RMSEA = .063, providing support for its inclusion in the study. In addition, as detailed in Results and Table 2, the central modeling involving negative motivation and engagement (that comprised independent variables and covariates) yielded satisfactory fit, $\chi^2(31) = 599$, CFI = .90, RMSEA = .038. Also, as further detailed in Results and Table 2, the statistically significant paths to negative motivation and engagement involved the same independent variables that significantly predicted (inversely) the reliable positive motivation and engagement factor. This suggests its marginal reliability did not disproportionately sway what would be predicted given the complementary findings for its reliable counterpart (positive motivation and engagement). Finally, the conduct of latent variable modeling in our study was a further way to help circumvent some measurement issues duc to unreliability. Taken together, we recognize that negative motivation and engagement presents marginal reliability (and urge the reader to interpret relevant parameters accordingly), but feel there is sufficient empirical support for its inclusion to offer a more rounded perspective on test-relevant motivation and engagement.

Subjective test experience. After completing all NAPLAN test items and test-relevant motivation and engagement measures, years 7 and 9 students were asked a series of seven questions exploring their subjective experience of the online test. Because of test length and burden considerations, the younger students (those in years 3 and 5) were not asked these additional questions. Whereas the motivation and engagement items (above) asked about issues relevant to “classic” motivation and engagement dimensions, the subjective test experience items were designed to tap into respondents’ reflections on how their personal experience of key facets of this test compared with other mathematics tests they had done before. Accordingly, questions addressed the *ease of test items*, the extent to which it *matched their ability*, their *interest in test items*, *item comprehension,* and their *perceived perfor-*

Table 2

Multi-Level SEM of Test Condition, Interactions, and Covariates Predicting Achievement, Motivation and Engagement, and Subjective Test Experience (N = 12,736; Years 3, 5, 7, and 9): Standardized Beta Coefficients, Standard Errors, and Model Fit

			Positive	Negative	
	Achievement	Achievement	motivation and	motivation and	Subjective test
	score	error	engagement	engagement	experience
	(Model 1)	(Model 2)	(Model 3)	(Model 4)	(Model 5)
Level	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)	β (*SE*)
Student-level (Level 1)					
Condition (fixed, adaptive – see note)	.02 (.01)	−.31 (.01)***	.01 (.01)	−.02 (.01)	−.03 (.04)
Gender (M, FM)	−.04 (.01)***	−.05 (.01)***	−.02 (.01)*	.10 (.01)***	−.15 (.02)***
Year-level	.01 (.01)	−.13 (.01)***	−.35 (.01)***	.12 (.02)***	−.35 (.03)***
Condition × Gender	.01 (.01)	.01 (.01)	.01 (.01)	−.01 (.01)	.04 (.01)**
Condition × Year-level	.04 (.01)***	−.02 (.01)	.02 (.01)*	−.03 (.01)**	.09 (.03)**
R square (explained variance)	.01 (.001)**	.12 (.01)***	.13 (.01)***	.03 (.01)***	.16 (.02)****
School-level (Level 2)					
Socio/Educ. advantage	.76 (.04)***	.41 (.10)***	.47 (.08)***	−.77 (.07)***	.60 (.09)***
Location	.09 (.06)	.11 (.11)	.05 (.10)	−.16 (.09)	−.06 (.11)
Structure (elem or sec, combined)	−.03 (.05)	.11 (.09)	.06 (.08)	.03 (.07)	−.10 (.07)
Size	.09 (.06)	.15 (.10)	.05 (.09)	−.13 (.08)	.16 (.10)
R square (explained variance)	.57 (.05)***	.22 (.08)***	.23 (.07)***	.57 (.08)***	.51 (.08)***
Model fit	$\chi^2(25) = 415$, CFI = 1.00, RMSEA < .001	$\chi^2(25) = 1505$, CFI = 1.00, RMSEA < .001	$\chi^2(63) = 1261$, CFI = .94, RMSEA = .039	$\chi^2(31) = 599$, CFI = .90, RMSEA = .038	$\chi^2(82) = 1120$, CFI = .90, RMSEA = .032

Note. Subsidiary tests for Condition effect sizes indicated the following: Achievement Score, d = .05; Achievement Error, d = .56; Positive Motivation and Engagement, d = .01; Negative Motivation and Engagement, d = .02; Subjective Test Experience, d = .15.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

mance. They also reported on the extent to which they saw items as *out of order* and any *item confusion*. In total, seven such items were administered (with the stem: "Compared to maths tests I have done before . . .," as follows: "This test was easy for me," "This test was suited to my ability," "I was interested in most of the questions in this test," "I understood most of the questions in this test," "I did well in this test," "The questions seemed to be random, or out of order," and "This test was unclear and confusing"). All items were rated on a 1 (*Strongly Disagree*) to 7 (*Strongly Agree*) scale. After reversing the two negatively worded items, using M*plus*, a latent subjective test experience factor was formed using the seven indicators (M = 4.40, SD = .93, Cronbach's alpha = .74, mean factor loading = .55). In variance components analysis (null model), the ICC (proportion of school-level variance) for subjective test experience was .04. Means and *SD*s for adaptive and fixed conditions are shown in Table 1.

Predictors and covariates. Student-level (Level 1) predictors and covariates comprised test condition (0 = fixed; 1 = adaptive), gender (0 = male, 1 = female), year-group (years 3, 5, 7, 9), Condition × Gender interaction, and Condition × Year-Group interaction (main effects were zero-centered prior to creating interaction terms; Aiken & West, 1991). To assist model parsimony in representing main and interaction effects, grade was treated as a continuous variable. However, where a statistically significant grade effect emerged, post hoc tests differentiated the effect as a function of distinct year groups (e.g., Table 3). School-level (Level 2) covariates were school socioeducational advantage, or SEA (ICSEA Score; national M = 1,000), location (an ordinal urban, semiurban/regional, remote, and very remote scale), structure (elementary or secondary only; combined elementary and secondary), and size (enrolment numbers).

Analyses

For the present investigation, the data were conceptualized as a two-level model, consisting of student at the first level (Level 1, L1) and school at the second level (Level 2, L2). The multilevel analyses (for details see Goldstein, 2003; Raudenbush & Bryk, 2002) were conducted using M*plus* version 7.31 (Muthén & Muthén, 2015). Missing survey data were imputed using the *mice* (Multivariate Imputation by Chained Equations) package in the *R* statistical software environment (Van Buuren & Groothuis-Oudshoorn, 2011). Full details of this imputation are provided in Supplementary Materials. For multilevel modeling in M*plus*, the comparative fit index (CFI) and root mean square error of approximation (RMSEA) were used as the benchmarks for fit. For CFI, a value of .90 or above is considered an acceptable fit (McDonald & Marsh, 1990). For RMSEA and SRMR, values less than or equal to .08 are deemed satisfactory fit (Marsh, Balla, & Hau, 1996; Schumacker & Lomax, 2010). Consistent with Figure 1, analyses comprised multilevel structural equation modeling (random intercept, fixed slope). At the student level (L1), predictors of achievement, latent motivation and engagement factors, and a latent subjective test experience factor were test condition (fixed order; adaptive), gender, year group, Condition × Gender interaction, and Condition × Year-Group interaction. At the school level (L2), predictors of achievement, latent motivation and engagement factors, and a latent subjective test experience factor were school SEA, school location, school structure, and school size. Thus, Level 1 independent variables predict student-level achievement, motivation and engagement, and subjective test experience, while Level 2 independent variables predict school-level achievement, motivation and engagement, and subjective test experience. We also make the point that although the analytic design is complex, the multilevel modeling and the multivariate set-up are intended to operate as appropriate controls for variance in order to better estimate unique effects/variance attributable to test condition (and its interactions).

Our focus is on the main effect of test condition and any interaction effects involving test condition. Although we emphasize interaction effects over main effects when both are statistically significant, we do also draw attention to the main effect in such cases. This is because there can be occasions when the treatment group's mean is always higher or lower than the comparison group's, signaling the fact that the treatment group not only changed more, but started higher or lower (Sweet & Grace-Martin, 2012). Statistically significant interaction terms were further pursued by SEM such that for each level of the relevant covariate (e.g., year-level), the effect of test condition was investigated. Thus, for example, for any statistically significant Condition × Year-Group interactions, SEM was conducted to examine the

Table 3
Post Hoc SEMs of Test Condition × Year-Level and Test Condition × Gender Interactions Predicting Achievement, Motivation and Engagement, and Subjective Test Experience: Standardized Beta Coefficients, Standard Errors, and Model Fit

Interaction	Achieve score β (*SE*)	Positive motivation & engagement β (*SE*)	Negative motivation & engagement β (*SE*)	Subjective test experience β (*SE*)
Year-level interactions				
Year 3 condition effect (fixed, adaptive)	−.01 (.02)	−.02 (.02)	.02 (.02)	NA
Year 5 condition effect (fixed, adaptive)	−.01 (.02)	−.02 (.02)	.01 (.02)	NA
Year 7 condition effect (fixed, adaptive)	.01 (.02)	.02 (.02)	−.05 (.03)	.03 (.02)
Year 9 condition effect (fixed, adaptive)	.12 (.02)***	.06 (.02)**	−.07 (.02)**	.12 (.02)***
Gender interactions				
Male condition effect (fixed, adaptive)	NS	NS	NS	.03 (.02)
Female condition effect (fixed, adaptive)	NS	NS	NS	.12 (.02)***

Note. NA = Year 3 and 5 were not administered these items to reduce the time burden on these younger students; NS = no statistically significant interaction in main analyses (Table 2) and so interactions are not pursued further.
$^{**}p < .01$. $^{***}p < .001$.

predictive path of test condition for each of years 3, 5, 7, and 9. Because test condition was the focus of the study and modeled at the student level (Level 1), no school-level (Level 2) interactions were investigated.

Results

Multilevel Modeling of Effects for Achievement

We examined the effects of test condition on students' achievement scores and achievement error rates (or, measurement precision). Table 2 shows results. For achievement scores (model fit: $\chi^2(25) = 415$, CFI = 1.00, RMSEA < .001), when controlling for student- and school-level covariates, there was no main effect for test condition on students' achievement scores (as per descriptive statistics in Table 1; see also effect size reported below Table 2, $d = 0.05$). There was, however, a statistically significant test Condition × Year-Level effect ($\beta = .04$, $p < .001$) such that year 9 students (relative to years 3, 5, and 7 students) scored better in the adaptive test condition than in the fixed order test condition; $\beta = .12$, $p < .001$ (see Table 3). Although not central to the study's focus on test condition, we note that student-level gender ($\beta = -.04$, $p < .001$; males achieving more highly) and school-level socioeducational advantage ($\beta = .76$, $p < .001$; higher SEA achieving more highly) yielded statistically significant effects for achievement.

In relation to achievement error rates (or, measurement precision; model fit: $\chi^2(25) = 1505$, CFI = 1.00, RMSEA < .001), when controlling for student- and school-level covariates, there was a statistically significant main effect for test condition such that there was less error (variability in correct responding from item-to-item) in the adaptive test condition than in the fixed order test condition, $\beta = -.31$, $p < .001$ (see Table 1 for descriptive statistics and Table 2 for all effects and model fit; see also effect size reported below Table 2, $d = 0.56$). No statistically significant interactions with test condition emerged. We also note that student-level gender ($\beta = -.05$, $p < .001$; males with higher error rates), year-level ($\beta = -.13$, $p < .001$; lower year-levels with higher error rates), and school-level socioeducational status ($\beta = .41$, $p < .001$; higher SES with higher error rates) yielded statistically significant effects for achievement error rates.

Multilevel Modeling of Effects for Motivation and Engagement

We examined the effects of test condition on students' test-relevant motivation and engagement. Table 2 shows results and model fit for positive and negative motivation and engagement factors (model fit: $\chi^2(63) = 1261$, CFI = .94, RMSEA = .039 and model fit: $\chi^2(31) = 599$, CFI = .90, RMSEA = .038, respectively). Findings demonstrate that when controlling for student- and school-level covariates, there was no statistically significant main effect for test condition on the motivation and engagement factors (also see descriptive statistics in Table 1; see also effect sizes reported below Table 2, $d = 0.01$ for positive motivation and engagement and $d = 0.02$ for negative motivation and engagement). There was, however, a statistically significant Condition × Year-Level interaction for positive motivation and engagement ($\beta = .02$, $p < .05$), with year 9 students (relative to years 3, 5, and 7 students) higher in positive motivation and engagement in the adaptive test condition than in the fixed order test condition ($\beta = .06$, $p < .01$). There was also a statistically significant Condition × Year-Level interaction for negative motivation and engagement ($\beta = -.03$, $p < .01$), with year 9 students (relative to years 3, 5, and 7 students) lower in negative motivation and engagement in the adaptive test condition than in the fixed order test condition ($\beta = -.07$, $p < .01$). Table 3 shows details.

Although not central to the substantive focus of the investigation, the following statistically significant covariate effects were found for positive motivation and engagement: student-level gender ($\beta = -.02$, $p < .01$; males with more positive motivation and engagement), year-level ($\beta = -.35$, $p < .001$; lower year-levels with more positive motivation and engagement), and school-level socioeducational advantage ($\beta = .47$, $p < .001$; higher SEA with more positive motivation and engagement). In addition, the following statistically significant covariate effects were found for negative motivation and engagement: student-level gender ($\beta = .10$, $p < .001$; females with more negative motivation and engagement), year-level ($\beta = .12$, $p < .001$; upper year-levels with more negative motivation and engagement), and school-level socioeducational advantage ($\beta = -.77$, $p < .001$; lower SEA with more negative motivation and engagement).

Although our focus is on the two latent ("global") motivation and engagement factors, results in Appendix B for the individual motivation and engagement indicators are illuminating. As shown in Table B1, when controlling for student- and school-level covariates, there was a statistically significant main effect for test condition on self-efficacy ($\beta = .02$, $p < .05$; higher for adaptive condition) and disengagement ($\beta = -.02$, $p < .05$; lower for adaptive condition) suggesting positive motivation and engagement effects for the adaptive condition. However, there was also a main effect for anxiety ($\beta = .02$, $p < .05$) suggesting higher anxiety for those in the adaptive condition. Statistically significant interactions with test condition also emerged and we describe and present these in Appendix B.

Multilevel Modeling of Effects for Subjective Test Experience

Year 7 and 9 students also reported on their subjective test experience (year 3 and 5 students were not administered these items to reduce the time burden on these younger students). Table 2 shows results (model fit: $\chi^2(82) = 1120$, CFI = .90, RMSEA = .032). Findings demonstrate that when controlling for student- and school-level covariates, there was no statistically significant main effect for test condition on the subjective test experience factor (in line with descriptive statistics in Table 1; see also effect size reported below Table 2, $d = 0.15$).

There was, however, a statistically significant interaction for subjective test experience between test condition and year-level ($\beta = .09$, $p < .01$) and also between test condition and gender ($\beta = .04$, $p < .01$). With regards to year-level, year 9 students (relative to year 7 students) reported more positive subjective test experience in the adaptive test condition than in the fixed order test condition ($\beta = .12$, $p < .001$). With regards to gender, females (relative to males) reported more positive subjective test experience in the adaptive test condition than in the fixed order test condition ($\beta = .12$, $p < .001$).

In addition to effects relevant to test condition, the following statistically significant covariate effects were found for subjective test experience: student-level gender ($\beta = -.15, p < .001$; males with more positive subjective experience), year-level ($\beta = -.35$, $p < .001$; lower year-levels with more positive subjective experience), and school-level socioeducational advantage ($\beta = .60, p < .001$; higher SEA with more positive subjective experience). Table 3 shows details.

Although our focus is on the latent ("global") subjective test experience factor, results in Appendix C and Table C1 for the individual subjective test experience indicators are also shown. These findings demonstrate that when controlling for student- and school-level covariates, there were no statistically significant main effects for test condition on any of the seven subjective experience indicators. Notably, however, statistically significant interactions with test condition did emerge for some of the individual subjective test experience indicators and we describe these in Appendix C.

Discussion

Although some prior work has attested to the merits of computer-adaptive testing, the research is not consistently positive (e.g., Colwell, 2013; Frey et al., 2009; Häusler & Sommer, 2008; Pitkin & Vispoel, 2001). Notwithstanding some critical issues to address (e.g., elevated anxiety - discussed below), the bulk of findings pointed to some positive computer-adaptive testing effects and apparently no negative ones. These results, alongside the logistic and administrative benefits associated with computer-adaptive testing (Stone & Davey, 2011), suggest it as a viable form of numeracy assessment with regards to test-relevant achievement, motivation, engagement, and subjective experience.

Findings of Particular Note to Research, Theory, and Practice

Our findings were consistent with claims that computer-adaptive testing better matches items to respondents, leading to less error and greater measurement precision within a test (Davey, 2011). As noted in the Introduction, less intratest error is a desirable measurement property in itself (Lazendic & Adams, 2014; Wise, 2015) and desirable for the test-taker's experience of the test (Lifelong Achievement Group & Martin, 2015). These findings also support conceptualizing relevant to adaptive testing. Item response theory (IRT)—a class of measurement models which provide the core methodological foundation for adaptive testing (Embretson & Reise, 2000; de Ayala, 2009; Kingsbury & Weiss, 1983; Stone & Davey, 2011)—emphasizes the importance of placing items and students on the same scale, facilitating the match between students and items (Thompson & Weiss, 2011).

Another major finding is the consistently positive effects of adaptive testing for the older (year 9) students in the sample. Relative to students in years 3, 5, and 7 (for whom there were no major differences between conventional and adaptive testing), the year 9 students demonstrated higher achievement levels, greater measurement precision, higher motivation and engagement, and more positive test experience. This stage of schooling (year 9) is known for its lower levels of academic motivation and engagement (Martin, 2007) and thus it is interesting that at the developmental period in which the risks of academic downturn are greater, adaptive computer testing yielded the opposite pattern. Stage-environment fit theory may be helpful in interpreting this finding. Stage-environment fit attributes declining motivation and engagement to mismatches between the needs of developing adolescents and the opportunities their high school environment provides (Eccles et al., 1993; Eccles & Roeser, 2009; Wang & Eccles, 2012). Computer-adaptive testing may offer better fit for different students at different levels of development. For year 9 students, the individualism that is a hallmark of adolescence (Blakemore, Burnett, & Dahl, 2010) may well suit the better match of items to students in the adaptive test condition. In addition, by year 9, students have sat many conventional paper-and-pencil tests and may be "conventional-test weary." Perhaps the role of technology in testing students at this developmental stage is somewhat refreshing and energizing for them, leading to greater engagement with the test and its demands. Taken together, with regard to the present findings for year 9 students, adaptive testing may be a useful consideration when testing students at a developmental stage in which fit with environment is poorest.

For all our dependent measures, school socioeducational advantage (SEA) explained the greatest variance. In some ways this is not surprising as large-scale and high-stakes testing has consistently shown SEA advantages (Colwell, 2013; Fritts & Marszalek, 2010; Perry & McConney, 2010). Students in high SEA schools tend to have higher intake characteristics, are better supported through their own (e.g., home) resources, and tend to be taught in better resourced environments (see Sirin, 2005 for a review). Thus, when considering each of the outcome measures, the role of SEA was more salient in impacting outcomes than the test condition. These findings once more underscore the importance of attending to socioeducational disparities in large-scale testing exercises.

Given the size of school-level SEA effects, it is unfortunate that we could not get information on the SEA of individual students (due to the test protocols governing this study). We can infer that student-level SEA effects might also be significant because a substantial upper-level effect often does correspond to a significant lower-level effect - but this is an empirical question we could not answer with our data. Given the role of school-level SEA in this study, then, future research would do well to include student-level SEA.

Although the study was focused on test-relevant motivation, engagement, and subjective experience as broader constructs, subsidiary analyses (see Appendixes B and C) with these constructs' lower-order indicators were illuminating. With regards to motivation and engagement, three specific factors were impacted by adaptive testing: self-efficacy, disengagement, and anxiety. Adaptive testing had positive effects for the former two, with students in the adaptive condition reporting higher test-relevant self-efficacy and lower test-relevant disengagement. The higher self-efficacy is consistent with Parshall, Spray, Kalohn, and Davey (2002) who suggest that better matched items promote self-efficacy through the task. Where items are better matched to students (and vice versa), a given student has a 50/50 chance of correctly responding; for many students this elevates opportunities for "success" and potentially by implication, self-efficacy. The lower disengagement is consistent with Way et al. (2010) who suggest that because content is delivered to a more appropriate level of difficulty, students may be more inclined to persist and complete the test.

Interestingly, however, students in the computer-adaptive testing condition reported higher anxiety—a finding in line with Colwell (2013) who observed that respondents are susceptible to the effects of test anxiety in computer-adaptive testing (see also Pitkin & Vispoel, 2001). Drawing on theories around personalized learning and growth orientations (e.g., Anderman, Gimbert, O'Connell, & Riegel, 2015; Dweck, 2006; Harris, 2011; Martin, 2015), we suggest that the better match between students and test items may generate a level of personalized challenge that increases students' arousal—detected via elevated anxiety in this study. Personal challenge is more aligned with arousal or emotionality than worry (see Liebert & Morris, 1967), which in turn influences academic outcomes in a positive manner (Cassady & Johnson, 2002). Likewise in the present study, elevated anxiety was not accompanied by a reduction in students' motivation and engagement; rather, students reported higher test-relevant self-efficacy and lower disengagement.

Practical Issues Going Forward

In the main, the present findings have demonstrated that there appear to be no salient negative effects of computer-adaptive testing in that fixed order testing was not superior to adaptive testing. In addition, on some measures there appear to be positive effects of adaptive tests when compared with a fixed test condition. Nevertheless, there are some critical issues important to consider in moving forward.

Because technology is costly, not all schools or school districts are able to provide the same amount or quality of technology. Access to computing may have implications for high-stakes computer-adaptive testing and the capacity to conduct fair testing across socioeconomic strata (Stone & Davey, 2011). Technological literacy is also a factor relevant to test performance, and so test administrators will need to guard against any potential advantage experienced by tech-literate and tech-advantaged schools and students (Stone & Davey, 2011).

Another set of critical issues concerns the need to acknowledge and consider threats to the validity of scores as well as interpretations that may affect adaptive testing. A common criticism of item-level adaptive testing is that all students are not completing the same items and even if the same items are answered, they may not be in the same order for all students (Way et al., 2010). Although tailoring is fundamental to adaptive testing and is an important means by which items presented to students can be reduced (whereas fixed order tests must present all items to cover all levels of ability), the possibility that different item or testlet presentation and ordering may impact the score for an individual test-taker cannot be discounted. Thus, test administrators are to be appropriately vigilant and nuanced when interpreting scores (Way et al., 2010). The stage-adaptive design (i.e., MAT) implemented in this study, and in future similar such tests (e.g., NAPLAN), offer advantages over item-level adaptive tests in relation to the control of item exposure, item order, and overall test content and coverage (Hendrickson, 2007; Zheng, Nozawa, Gao, & Chang, 2012).

Way et al. (2010) also point out there have been long and hard-fought battles to have students with special needs and English-language learners included in the curriculum and in appropriate and important assessment exercises. We believe that adaptive testing (that better matches items to students) may be well suited to such students in that it is aspirationally underpinned by scaffolding students to items that are personally and optimally challenging. This advice is consistent with well-established growth, personal best, and individualized approaches to student achievement, motivation, and engagement (e.g., Anderman et al., 2015; Dweck, 2006; Harris, 2011; Martin, 2015). In addition, it is important to be mindful that the IRT-based technique has been criticized in its strong assumption of unidimensionality that has pupils placed onto a common scale. This risks downplaying multidimensional views of students' capacities. There are thus inherent tensions relevant to computer-adaptive testing that are relevant to its implementation and interpretation.

In moving forward with computer-adaptive testing, we recommend consideration of potential vulnerabilities that may be triggered by this assessment method. We identified test-relevant anxiety as one such factor. Although we suggest our study triggered arousal more than anxiety (Cassady & Johnson, 2002; Liebert & Morris, 1967)—because it was accompanied by higher self-efficacy and lower disengagement—we advise caution. Colwell (2013) reported that test-takers who are allowed to review their answers in an adaptive condition tend to report lower anxiety than test-takers who are not allowed review opportunities. In other work, test-takers who are allowed to choose the level of item difficulty also demonstrate less anxiety (Shermis, Mzunara, & Bublitz, 2001).

Limitations and Future Directions

There are some limitations important to consider when interpreting our findings and which have implications for future research in adaptive testing. First, as descriptive statistics in Table 1 and modeling in Table 2 show, the statistically significant effects for adaptive testing were not large (the effects of school SEA, for example, were much larger). However, we suggest the fact they are statistically significantly positive at all is an important insight as it demonstrates there are apparently few negative effects using an adaptive approach to testing. Second, although the study comprised objective achievement measures, there were also self-report data that have known limitations. These data are susceptible to response bias (Wise & Ma, 2012) and findings must be interpreted with this in mind. In addition, we opted to focus on standard error as a measure of computer-adaptive testing efficiency, but its efficiency can also be demonstrated algebraically (Hambleton, 1989; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1971a). Third, there is the possibility that test-takers who are not motivated or engaged during the test will also not be motivated or engaged to respond accurately to survey items presented at the end of the test (Wise & Ma, 2012). Fourth, because our motivation, engagement, and subjective experience measures were taken at the end of the test, they were not sensitive to changes through the test. Methods to capture such data in future testing would be helpful. Fifth, as noted earlier, our data were from a nationally based assessment exercise administered by a central authority. The available data were subject to restrictions and thus limited student-level data were included in analyses. Thus, for example, we could not control for achievement on prior numeracy tests and this is important to address in future studies. Similarly, although we assigned students to adaptive and fixed order conditions using a "round-robin" approach (described in Procedure), we did not have much data by which to establish preexisting equivalence between the two conditions. Sixth, adaptive testing has distinct implications for students with special needs (Stone & Davey, 2011) as well as for students who may benefit from significantly advanced items (Ediger, 2007). There may thus be unique motivation and engagement effects specific to these groups but the data available could not address this. This is another issue for future research.

A seventh point concerns the study's nationally based assessment approach that was focused on administering the full instrumentation to participants; thus, it was only possible to administer the short form of the Motivation and Engagement Scale. Although this Scale's score has demonstrated reliability and valid inferences in prior research (see Liem & Martin, 2012) and we operationalized motivation and engagement via higher order latent factors using the short from indicators (including for the subjective test experience factor), findings are to be interpreted accordingly. Eighth, our study was located in the numeracy domain and so further research is needed to assess the generality of current findings to the literacy domain (the other area assessed under NAPLAN). Ninth, to control for the role of the computer in our assessment exercise, we did not include paper-and-pencil testing as a condition. Further work might seek to consider how computer-adaptive testing compares with fixed paper-and-pencil testing.

We also point out that different mathematics subdomains were aggregated in our analyses and that ideas and tasks in these subdomains do change between year 3 and year 9. While recognizing that aggregation loses some nuances available through analyses within subdomain as a function of year-level, we sought to guard against this in numerous ways: we separated out results as a function of year-level and gender; we conducted analyses based on subdomains that were common across years 3–9; we ran Rasch models to ascertain fit statistics for each year-level; we ensured comparable reliability for achievement scores for each year group; and we standardized achievement scores by year-level. Another point relevant to comparability is that standardizing within year can make it difficult to interpret across-year interactions. We thus note these interactions as statistically significant but advise some caution in interpreting their specific size. Finally, the issues under focus here lend themselves to qualitative research allowing participants to articulate in rich and nuanced ways how the adaptive condition impacts their experience of the test and their test perceptions. Preliminary work along these lines has been conducted using cognitive interviewing (Lowrie & Logan, 2013) and there is now a need to expand qualitative enquiry into the realm of test-relevant motivation and engagement.

Conclusion

Although there is rather consistent evidence showing computer-adaptive testing increases the precision of student achievement estimates, there has been far less research with regard to its effects on students' test-relevant motivation, engagement, and subjective experience. The present findings have provided much needed detail on the role of computer-adaptive testing with respect to these important test-relevant outcomes. As expected, students in the computer-adaptive testing condition generated lower achievement error rates (i.e., higher measurement precision). On the other dependent measures (motivation, engagement, subjective test experience), there were no significant advantages or disadvantages of computer-adaptive testing as a main effect, however significant positive computer-adaptive test effects did emerge as a function of year-level and gender. We suggest findings hold implications for researchers and test administrators as they seek ways to promote more positive experiences for students as they participate in large-scale computer-based and online testing.

References

Abedi, J., Leon, S., & Kao, J. (2007). *Examining differential distractor functioning in reading assessments for students with disabilities*. Minneapolis, MN: University of Minnesota, Partnership for Accessible Reading Assessment.

Adams, J. R., & Lazendic, G. (2013). *Observations on the feasibility of a multistage test design for NAPLAN*. Unpublished Tech. Rep. No.

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31,* 162–172. http://dx.doi.org/10.1016/j.stueduc.2005.05.008

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. London, UK: Sage.

Allison, P. D. (2015). *Predictive mean matching*. Retrieved from http://statisticalhorizons.com/predictive-mean-matching

Alston, M. M., & Kent, J. (2003). Educational access for Australia's rural young people: A case of social exclusion. *Australian Journal of Education, 47,* 5–17. http://dx.doi.org/10.1177/000494410304700102

Anderman, E. M., Gimbert, B., O'Connell, A. A., & Riegel, L. (2015). Approaches to academic growth assessment. *British Journal of Educational Psychology, 85,* 138–153. http://dx.doi.org/10.1111/bjep.12053

Australian Curriculum, Assessment, and Reporting Authority (ACARA). (2014a). *NAPLAN: National Protocols for Test Administration 2014*. Sydney, Australia: Author.

Australian Curriculum, Assessment, and Reporting Authority (ACARA). (2013). *Fact sheet: Moving NAPLAN online—ACARA research and findings*. Sydney, Australia: Author.

Australian Curriculum, Assessment, and Reporting Authority (ACARA). (2014b). *National Assessment and Surveys Online Program: Tailored test design 2013 research study summary*. Sydney, Australia: Author.

Australian Curriculum, Assessment, and Reporting Authority (ACARA). (2014c). *Tailored test design study 2013: Summary research report*. Sydney, Australia: Author.

Australian Human Rights Commission. (2000). *Education access: National inquiry into rural and remote education*. Sydney, Australia: Human Rights and Equal Opportunity Commission.

Blakemore, S. J., Burnett, S., & Dahl, R. E. (2010). The role of puberty in the developing adolescent brain. *Human Brain Mapping, 31,* 926–933. http://dx.doi.org/10.1002/hbm.21052

Booth, M. Z., & Gerard, J. M. (2014). Adolescents' stage-environment fit in middle and high school: The relationship between students' perceptions of their schools and themselves. *Youth & Society, 46,* 735–755. http://dx.doi.org/10.1177/0044118X12451276

Bourke, L. (1997). Outlook of rural secondary students: A preliminary case study in North Queensland. *Youth Studies Australia, 16,* 11–16.

Breithaupt, K., & Hare, D. R. (2007). Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement, 67,* 5–20. http://dx.doi.org/10.1177/0013164406288162

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27,* 270–295. http://dx.doi.org/10.1006/ceps.2001.1094

Colwell, N. M. (2013). Test anxiety, computer-adaptive testing, and the common core. *Journal of Education and Training Studies, 1,* 50–60. http://dx.doi.org/10.11114/jets.v1i2.101

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York, NY: Harper & Row.

Cudeck, R. (1985). A structural comparison of conventional and adaptive versions of the ASVAB. *Multivariate Behavioral Research, 20,* 305–322. http://dx.doi.org/10.1207/s15327906mbr2003_5

Davey, T. (2011). *A guide to computer adaptive testing systems.* Washington, DC: Council of Chief State School Officers.

De Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: Guilford Press.

Dweck, C. S. (2006). *Mindset: The new psychology of success.* New York, NY: Random House.

Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., & Iver, D. M. (1993). Development during adolescence. The impact of stage-environment fit on young adolescents' experiences in schools and in families. *American Psychologist, 48,* 90–101. http://dx.doi.org/10.1037/0003-066X.48.2.90

Eccles, J. S., & Roeser, R. W. (2009). Schools, academic motivation, and stage-environment fit. In R. M. Lerner & L. Steinber (Eds.), *Handbook of adolescent psychology* (3rd ed., pp. 404–434). Hoboken, NJ: Wiley.

Ediger, M. (2007). *Curriculum organization.* New Delhi, India: Discovery Publishing House.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Fredricks, J. A., & Eccles, J. S. (2002). Children's competence and value beliefs from childhood through adolescence: Growth trajectories in two male-sex-typed domains. *Developmental Psychology, 38,* 519–533. http://dx.doi.org/10.1037/0012-1649.38.4.519

Frey, A., Hartig, J., & Moosbrugger, H. (2009). Effects of adaptive testing on test taking motivation. *Diagnostica, 55,* 20–28. http://dx.doi.org/10.1026/0012-1924.55.1.20

Fritts, B. E., & Marszalek, J. (2010). Computerized adaptive testing, anxiety levels, and gender differences. *Social Psychology of Education, 13,* 441–458. http://dx.doi.org/10.1007/s11218-010-9113-3

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., . . . Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59,* 361–368. http://dx.doi.org/10.1176/ps.2008.59.4.361

Gillen-O'Neel, C. G., & Fuligni, A. (2013). A longitudinal study of school belonging and academic motivation across high school. *Child Development, 84,* 678–692. http://dx.doi.org/10.1111/j.1467-8624.2012.01862.x

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London, UK: Hodder Arnold.

Green, J., Martin, A. J., & Marsh, H. W. (2007). Motivation and engagement in English, mathematics and science high school subjects: Towards an understanding of multidimensional domain specificity. *Learning and Individual Differences, 17,* 269–279. http://dx.doi.org/10.1016/j.lindif.2006.12.003

Gulliksen, H. (1950). *Theory of mental tests.* New York, NY: Wiley. http://dx.doi.org/10.1037/13240-000

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 147–200). New York, NY: Macmillan.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Harris, D. N. (2011). *Value-added measures in education.* Cambridge, UK: Harvard Educational Press.

Hattie, J. A. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London, UK: Routledge.

Häusler, J., & Sommer, M. (2008). The effect of success probability on test economy and self-confidence in computerized adaptive tests. *Psychology Science, 50,* 75–87.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26,* 44–52. http://dx.doi.org/10.1111/j.1745-3992.2007.00093.x

Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in children's self-competence and values: Gender and domain differences across grades one through twelve. *Child Development, 73,* 509–527. http://dx.doi.org/10.1111/1467-8624.00421

Johnstone, C. J., Altman, J., & Thurlow, M. (2006). *A state guide to the development of universally designed assessments.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Kingsbury, G. G., & Hauser, C. (2004). *Computerized adaptive testing.* Portland, OR: Northwest Evaluation Association.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257–283). New York, NY: Academic Press. http://dx.doi.org/10.1016/B978-0-12-742780-5.50024-X

Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education, 22,* 22–37. http://dx.doi.org/10.1080/08957340802558326

Lazendic, G., & Adams, J. R. (2014, April). *Multistage test design incorporating vertical scale.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Philadelphia, PA.

Lee, V. E., & Smith, J. B. (1997). High school size: Which works best and for whom? *Educational Evaluation and Policy Analysis, 19,* 205–227. http://dx.doi.org/10.3102/01623737019003205

Liebert, R. M., & Morris, L. W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports, 20,* 975–978. http://dx.doi.org/10.2466/pr0.1967.20.3.975

Liem, G. A., & Martin, A. J. (2012). The Motivation and Engagement Scale: Theoretical framework, psychometric properties, and applied yields. *Australian Psychologist, 47,* 3–13. http://dx.doi.org/10.1111/j.1742-9544.2011.00049.x

Lifelong Achievement Group & Martin A. J. (2015). *Online NAPLAN testing and student motivation: Exploring adaptive and fixed test formats.* Unpublished Report to ACARA. Sydney, Australia: Lifelong Achievement Group.

Little, R. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics, 6,* 287–296.

Lord, F. M. (1971a). A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement, 31,* 805–813. http://dx.doi.org/10.1177/001316447103100401

Lord, F. M. (1971b). A theoretical study of two-stage testing. *Psychometrika, 36,* 227–242. http://dx.doi.org/10.1007/BF02297844

Lord, F. M. (1971c). Tailored testing, an approximation of stochastic approximation. *Journal of the American Statistical Association, 66,* 707–711. http://dx.doi.org/10.1080/01621459.1971.10482333

Lowrie, T., & Logan, T. (2013). *NAPLAN online: Trial of tailored test design—Numeracy cognitive interviews.* Wagga Wagga, Australia: Research Institute for Professional Practice, Learning and Education, Charles Sturt University.

Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education, 19,* 189–202. http://dx.doi.org/10.1207/s15324818ame1903_2

Mårdberg, B., & Carlstedt, B. (1998). Swedish Enlistment Battery (SEB): Construct validity and latent variable estimation of cognitive abilities by the CAT-SEB. *International Journal of Selection and Assessment, 6,* 107–114. http://dx.doi.org/10.1111/1468-2389.00079

Marsh, H. W., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical processes. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling techniques*. Hillsdale, NJ: Erlbaum.

Marsh, H. W., Martin, A. J., & Cheng, J. (2008). A multilevel perspective on gender in classroom motivation and climate: Potential benefits of male teachers for boys? *Journal of Educational Psychology, 100,* 78–95. http://dx.doi.org/10.1037/0022-0663.100.1.78

Martin, A. J. (2001). The Student Motivation Scale: A tool for measuring and enhancing motivation. *Australian Journal of Guidance & Counselling, 11,* 1–20. http://dx.doi.org/10.1017/S1037291100004301

Martin, A. J. (2007). Examining a multidimensional model of student motivation and engagement using a construct validation approach. *British Journal of Educational Psychology, 77,* 413–440. http://dx.doi.org/10.1348/000709906X118036

Martin, A. J. (2009). Motivation and engagement across the academic lifespan: A developmental construct validity study of elementary school, high school, and university/college students. *Educational and Psychological Measurement, 69,* 794–824. http://dx.doi.org/10.1177/0013164409332214

Martin, A. J. (2014). *The Motivation and Engagement Scale* (14th ed.). Sydney, Australia: Lifelong Achievement Group. Retrieved from http://www.lifelongachievement.com

Martin, A. J. (2015). Growth approaches to academic development: Research into academic trajectories and growth assessment, goals, and mindsets. *British Journal of Educational Psychology, 85,* 133–137. http://dx.doi.org/10.1111/bjep.12071

Martin, A. J., Anderson, J., Bobis, J., Way, J., & Vellar, R. (2012). Switching on and switching off in mathematics: An ecological study of future intent and disengagement amongst middle school students. *Journal of Educational Psychology, 104,* 1–18. http://dx.doi.org/10.1037/a0025988

Martin, A. J., Mansour, M., Anderson, M., Gibson, R., Liem, G. A. D., & Sudmalis, D. (2013). The role of arts participation in students' academic and non-academic outcomes: A longitudinal study of school, home, and community factors. *Journal of Educational Psychology, 105,* 709–727. http://dx.doi.org/10.1037/a0032795

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin, 107,* 247–255. http://dx.doi.org/10.1037/0033-2909.107.2.247

Ministerial Council on Education. Employment, Training and Youth Affairs (MCEETYA). (2006). *Statements of Learning for Mathematics.* Retrieved from http://www.curriculum.edu.au/verve/_resources/SOL_Maths_Copyright_update2008.pdf

Minnema, J., Thurlow, M., Bielinski, J., & Scott, J. (2000). *Past and present understandings of out-of-level testing: A research synthesis (Out-of-Level Testing Report 1).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Moreno, K. E., & Segall, O. D. (1997). Reliability and construct validity of CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 169–180). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10244-018

Mulvenon, S. W., Stegman, C. E., & Ritter, G. (2005). Test anxiety: A multifaceted study on the perceptions of teachers, principals, counselors, students, and parents. *International Journal of Testing, 5,* 37–61. http://dx.doi.org/10.1207/s15327574ijt0501_4

Muthén, L. K., & Muthén, B. O. (2015). *Mplus user's guide.* Los Angeles, CA: Author.

Newman, M., Garrett, Z., Elbourne, D., Bradley, S., Noden, P., Taylor, J., & West, A. (2006). Does secondary school size make a difference? A systematic review. *Educational Research Review, 1,* 41–60. http://dx.doi.org/10.1016/j.edurev.2006.03.001

Olson, L. (2003). Legal twists, digital turns: Computerized testing feels the impact of "No Child Left Behind." *Education Week, 12,* 11–14, 16.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. C. (2002). *Practical considerations in computer-based testing.* New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4613-0083-0

Perry, L., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers College Record, 112,* 1137–1162.

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95,* 667–686. http://dx.doi.org/10.1037/0022-0663.95.4.667

Pitkin, A. K., & Vispoel, W. P. (2001). Differences between self-adapted and computerized adaptive tests: A meta-analysis. *Journal of Educational Measurement, 38,* 235–247. http://dx.doi.org/10.1111/j.1745-3984.2001.tb01125.x

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press. (Original work published 1960)

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reeve, J. (2016). A grand theory of motivation: Why not? *Motivation and Emotion, 40,* 31–35. http://dx.doi.org/10.1007/s11031-015-9538-2

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997). *Computerized adaptive testing. From inquiry to operation.* Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10244-000

Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling.* New York, NY: Routledge.

Shermis, M. D., Mzunara, H. R., & Bublitz, S. T. (2001). On test and computer anxiety: Test performance under CAT and SAT conditions. *Journal of Educational Computing Research, 24,* 57–75. http://dx.doi.org/10.2190/4809-38LD-EEUF-6GG7

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75,* 417–453. http://dx.doi.org/10.3102/00346543075003417

Smarter Balanced Assessment Consortia. (2015). *Smarter Balanced Assessment Consortium: 2013–14 Tech. Rep. No.* Retrieved from http://www.smarterbalanced.org/wp-content/uploads/2015/08/2013-14_Technical_Report.pdf

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology, 8,* 33. http://dx.doi.org/10.1186/1471-2288-8-33

Stone, E., & Davey, T. (2011). *Computer-adaptive testing for students with disabilities.* Princeton, NJ: Educational Testing Service.

Sweet, S. A., & Grace-Martin, K. (2012). *Data analysis with SPSS.* New York, NY: Pearson.

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16,* 1–9.

Thurlow, M., Lazarus, S. S., Albus, D., & Hodgson, J. (2010). *Computer-based testing: Practices and considerations (Synthesis Report No. 78).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Triplett, C. F., Barksdale, M. A., & Leftwich, P. (2003). High stakes for whom? *Journal of Research in Education, 13,* 15–21.

Van Buuren, S. (2012). *Flexible imputation of missing data.* Boca Raton, FL: Chapman & Hall/CRC. http://dx.doi.org/10.1201/b11826

Van Buuren, S., & Groothuis-Oudshoorn, C. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45,* 1–67.

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures. *Applied Measurement in Education, 7,* 53–79. http://dx.doi.org/10.1207/s15324818ame0701_5

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Mislevy, R., & Thissen, D. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale, NJ: Erlbaum.

Wang, M. T., & Eccles, J. S. (2012). Adolescent behavioral, emotional, and cognitive engagement trajectories in school and their differential relations to educational success. *Journal of Research on Adolescence, 22,* 31–39. http://dx.doi.org/10.1111/j.1532-7795.2011.00753.x

Way, W. D., Twing, J. S., Camara, W., Sweeney, K., Lazer, S., & Maeo, J. (2010). *Some considerations related to the use of adaptive testing for the common core assessments.* Princeton, NJ: Educational Testing Service.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement. *Measurement and Evaluation in Counseling and Development, 37,* 70–84.

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education, 28,* 237–252. http://dx.doi.org/10.1080/08957347.2015.1042155

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method.* Paper presented at the 2012 annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Zheng, Y., Nozawa, Y., Gao, X., & Chang, H. H. (2012). *Multistage adaptive testing for a large-scale classification test.* ACT Research Report Series, 2012 (6). Retrieved from http://media.act.org/documents/ACT_RR2012-6.pdf

Appendix A

Distribution of Sample as a Function of Australian Test Administration Authority (TAA)

	Test Administration Authority (TAA)								
Measure	ACT	NSW	NT	Qld.	SA	Tas.	Vic.	WA	Total
Number of schools in the study	8	82	7	29	14	9	43	39	231
Percentage of schools in the study	3.5%	35.5%	3.0%	12.6%	6.0%	3.9%	18.6%	16.9%	100%
Percentage of schools in Australia	1.4%	29.8%	1.8%	20.0%	9.3%	2.9%	22.1%	12.7%	100%

Note. ACT = Australian Capital Territory; NSW = New South Wales; NT = Northern Territory; Qld = Queensland; SA = South Australia; Tas = Tasmania; Vic = Victoria; WA = Western Australia.

(*Appendices continue*)

Appendix B

Effects for Individual Motivation and Engagement Indicators

Although our study's focus is on the latent motivation and engagement factors (shown in Tables 2 and 3), for completeness we here present findings for the individual motivation and engagement indicators. As shown in Table B1, when controlling for student- and school-level covariates, there was a statistically significant main effect for test condition on self-efficacy ($\beta = .02$, $p < .05$; higher for adaptive condition) and disengagement ($\beta = -.02$, $p < .05$; lower for adaptive condition) suggesting positive motivation and engagement effects for the adaptive condition. However, there was also a main effect for anxiety ($\beta = .02$, $p < .05$) suggesting higher anxiety for those in the adaptive condition.

There were statistically significant interaction effects for individual motivation and engagement indicators. There was a statistically significant Condition × Gender interaction for uncertain control ($\beta = -.02$, $p < .05$), with females (relative to males) scoring lower in uncertain control in the adaptive test condition than in the fixed order test condition ($\beta = -.03$, $p < .05$). Three statistically significant Condition × Year-Group interactions emerged; one for self-efficacy ($\beta = .03$, $p < .001$), one for anxiety ($\beta = -.02$, $p < .05$), and another for disengagement ($\beta = -.03$, $p < .001$). For self-efficacy, year 7 students ($\beta = .04$, $p < .05$) and to a greater extent year 9 students ($\beta = .09$, $p < .001$) (relative to years 3 and 5 students) reported higher self-efficacy in the adaptive test condition than in the fixed order test condition. For anxiety, year 9 students ($\beta = -.04$, $p < .05$) (relative to years 3, 5, and 7 students) reported lower anxiety in the adaptive test condition than in the fixed order test condition. For disengagement, year 7 students ($\beta = -.05$, $p < .05$) and to a greater extent year 9 students ($\beta = -.07$, $p < .01$) (relative to years 3 and 5 students) reported lower disengagement in the adaptive test condition than in the fixed order test condition.

Table B1
Multi-Level SEM of Test Condition, Interactions, and Covariates Predicting Specific Indicators of Motivation and Engagement (N = 12,736; Years 3, 5, 7, and 9): Standardized Beta Coefficients and Standard Errors

	Positive motivation and engagement						Negative motivation and engagement			
Level	Self-efficacy	Valuing	Mastery orient	Planning	Task management	Persistence	Anxiety	Uncertain control	Self-handicapping	Disengagement
Student-level (Level 1)										
Condition (fixed, adaptive)	$.02^{*}$	$-.01$	.01	.01	.01	$-.01$	$.02^{*}$	$-.01$	$-.01$	$-.02^{*}$
Gender (M, FM)	$-.10^{***}$	.01	$-.03^{**}$	.01	$.02^{*}$	$.03^{**}$	$.07^{***}$	$.05^{***}$	$-.01$	$.08^{***}$
Year-level	$-.29^{***}$	$-.39^{***}$	$-.21^{***}$	$-.15^{***}$	$-.23^{***}$	$-.09^{***}$	$-.14^{***}$	.01	$.07^{***}$	$.13^{***}$
Condition × Gender	.01	.01	.01	$-.01$	.01	.01	.01	$-.02^{*}$	$-.01$	.01
Condition × Year-level	$.03^{***}$	.01	.02	.01	.01	.01	$-.02^{*}$	$-.01$	$-.01$	$-.03^{***}$
School−level (Level 2)										
Socio/Educ. advantage	$.40^{***}$	$-.23^{*}$	$.51^{***}$	$.23^{*}$	$.42^{***}$	$.56^{***}$	$-.67^{***}$	$-.83^{***}$	$-.57^{***}$	$-.65^{***}$
Location	.07	$-.14$	.04	$-.07$	.10	.13	$-.22^{*}$	$-.25^{*}$	$-.11$	$-.05$
Structure (elem or sec, combined)	$-.04$	.05	.01	$-.02$	.07	$-.05$	.12	.07	.01	.02
Size	.06	$-.10$	.04	.03	.07	.01	$-.19^{*}$	$-.09$	$-.11$	$-.11$

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

(*Appendices continue*)

Appendix C

Effects for Individual Subjective Test Experience Indicators

Although our study's focus is on the latent subjective test experience factor (shown in Tables 2 and 3), for completeness we here present findings for the individual subjective test experience indicators. As shown in Table C1, when controlling for student- and school-level covariates, there were no statistically significant main effects for test condition on any of the seven subjective experience measures.

Notably, however, there were statistically significant interaction effects for individual subjective test experience indicators. Three statistically significant Condition × Gender effects emerged: for item comprehension (β = .03, $p < .05$), perception of positive performance (β = .04, $p < .01$), and belief the test was confusing (β = −.04, $p < .01$). Female students (relative to male students) evinced higher item comprehension means in the adaptive test condition than in the fixed order test condition (β = .10, $p < .001$); higher perceived performance means in the adaptive test condition than in the fixed order test condition (β = .09, $p < .001$); and, lower means on test confusion in the adaptive test condition than in the fixed order test condition (β = −.08, $p < .001$).

Additionally, five statistically significant Condition × Year-Group interaction effects emerged for individual subjective test experience indicators: perceived ease of test (β = .06, $p < .05$), belief the test suited ability (β = .07, $p < .05$), item comprehension (β = .07, $p < .05$), perceived performance (β = .06, $p < .05$), and belief the test was confusing (β = −.09, $p < .01$). Year 9 students (relative to year 7 students) reported higher test ease in the adaptive test condition than in the fixed order test condition (β = .06, $p < .01$); year 7 students (β = .05, $p < .05$) and to a greater extent year 9 students (β = .11, $p < .001$) believed the test better suited their ability in the adaptive test condition than in the fixed order test condition; year 9 students (relative to year 7 students) reported greater item comprehension in the adaptive test condition than in the fixed order test condition (β = .10, $p < .001$); year 9 students (relative to year 7 students) reported higher perceived performance in the adaptive test condition than in the fixed order test condition (β = .09, $p < .001$); and, year 9 students (relative to year 7 students) were less likely to see the test as confusing in the adaptive test condition than in the fixed order test condition (β = −.09, $p < .001$).

Table C1
Multi-Level SEM of Test Condition, Interactions, and Covariates Predicting Specific Indicators of Subjective Test Experience (N = 5,382; Years 7 and 9): Standardized Beta Coefficients and Standard Errors

	Positive subjective test experience					Negative subjective test experience	
Level	Test easy for me	Test suited my ability	Interested in items	Understood items	Did well	Items random or out of order	Test confusing
Student-Level (Level 1)							
Condition (Fixed, Adaptive)	−.04	.01	.04	−.02	−.02	−.01	.06
Gender (M, FM)	−.12***	−.04**	−.03*	−.10***	−.16***	−.01	.05***
Year-level	−.07*	−.23***	−.33***	−.24***	−.35***	.15***	.19***
Condition × Gender	.02	.01	.01	.03*	.04**	.01	−.04**
Condition × Year-level	.06*	.07*	−.03	.07*	.06*	−.04	−.09**
School-Level (Level 2)							
Socio-Educ. Advantage	.60***	.42**	−.08	.62***	.43**	−.14	−.47***
Location	−.09	−.30	−.25	−.08	−.16	−.15	.09
Structure (Elem or Sec, Combined)	−.07	−.07	.07	−.07	−.02	.26*	.11
Size	.10	−.10	.03	.14	.02	−.15	−.20

Note. Year 3 and 5 were not administered these items to reduce the time burden on these younger students.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Received April 17, 2016
Revision received February 27, 2017
Accepted March 13, 2017 ■

Journal of Educational Psychology
2018, Vol. 110, No. 1, 46–57

0022-0663/18/$12.00 http://dx.doi.org/10.1037/edu0000196

When Students Doubt Their Teachers' Diagnostic Competence: Moderation in the Internal/External Frame of Reference Model

Friederike Zimmermann and Jens Möller
Kiel University

Olaf Köller
Leibniz Institute for Science and Mathematics Education (IPN)

The internal/external frame of reference model (I/E model) posits that individuals' achievement-related self-concepts are formed through social comparisons (e.g., self vs. peers) within academic domains and dimensional comparisons (e.g., math vs. verbal) between distinct domains. A large body of research has supported the theorized pattern of positive within-domain and negative cross-domain effects of achievement on self-concept. However, research on moderators of these effects has been scarce. In this article, we report results from 2 samples of 7th graders ($N_{\text{sample1}} = 1{,}045$ and $N_{\text{sample2}} = 1{,}966$) in which we investigated whether students' perceptions of teachers' domain-specific competence in judging students' achievement moderated the relation between teacher-assigned grades and students' self-concepts in the I/E model. Structural equation modeling was used to estimate the I/E model, which also included teachers' perceived diagnostic competence in each domain (i.e., math and verbal) and 4 latent interaction variables derived from that competence and from school grades in both domains. Both studies found support for the predictions of the I/E model. The cross-domain effect on math self-concept was consistently moderated in both studies: The lower students perceived their teachers' competence in diagnosing math achievement, the stronger the negative path from verbal grade to math self-concept. Furthermore, the results tentatively imply that the positive within-domain effect of verbal grade on verbal self-concept is stronger when students' perceptions of teachers' diagnostic competence in judging verbal achievement are higher. The results extend knowledge on the conditions of academic self-concept formation and provide valuable insights into possible ramifications of teachers' judgment accuracy.

Educational Impact and Implications Statement

This study suggests that students' beliefs in their competence partly depend on their beliefs in their teachers' ability to judge the students' achievement correctly. That means, for example, when students perceive their teachers' competence in diagnosing verbal achievement to be high, verbal grades may more strongly impact students' verbal competence belief. Our results show in particular that, when students perceive their teachers' competence in diagnosing math achievement to be low, students' math self-concept is influenced more strongly by other sources than math grades, namely, by contrasting with verbal achievement. The results point out the importance of teachers' diagnostic competence.

Keywords: academic self-concept, comparison processes, dimensional comparison, perceived accuracy of teachers' judgments, moderation

Students' perceptions of their own academic abilities are core elements of their identities and among the more important factors for their motivation and achievement (Valentine, DuBois, & Cooper, 2004). Self-concept features prominently in major theoretical accounts of motivational research, including expectancy-value theory (Eccles, 1983), self-determination theory (Deci & Ryan, 2002), and Pintrich's (2003) delineation of motivational science. Eccles (1983), for example, proposed that academic self-concept is influenced by prior scholastic failure and success and plays a central role in the development of motivation and academic choice

This article was published Online First March 23, 2017.

Friederike Zimmermann and Jens Möller, Institute for Psychology of Learning and Instruction, Kiel University; Olaf Köller, Department of Educational Research, Leibniz Institute for Science and Mathematics Education (IPN).

The research reported in this article is part of the project Development and Implementation of a School-to-Work Transition Concept for Schools Serving Disadvantaged Communities (Entwicklung und Implementierung eines neuen Konzeptes zur Eingliederung Jugendlicher in die Berufs- und Arbeitswelt in Schulen mit erhöhtem Förderbedarf [EIKA]), which is under the direction of Olaf Köller. The project was funded by the Senator for Education and Science of the Free Hanseatic City of Bremen and was cofinanced by the European Regional Development Fund (ERDF).

Correspondence concerning this article should be addressed to Friederike Zimmermann, Institute for Psychology of Learning and Instruction, Kiel University, Olshausenstrasse 75, 24118 Kiel, Germany. E-mail: zimmermann@psychologie.uni-kiel.de

behavior. High academic self-concept is indeed associated with positive scholastic outcomes such as motivation and achievement (e.g., Valentine et al., 2004).

Scholars from various theoretical approaches have emphasized that academic self-concept is affected by students' achievement, possibly through comparison processes (Marsh, 1986; Marsh & Craven, 2006; Möller & Marsh, 2013; Mussweiler, Rüter, & Epstude, 2006; Suls & Wheeler, 2000). Two of the more prominent in explaining the formation of self-concept are social and dimensional comparisons. Social comparisons use others as a standard for evaluating oneself, whereas dimensional comparisons involve using one's own achievements in other domains as the standard for evaluation (Biernat & Eidelman, 2007; Eccles, 2009; Festinger, 1954). In his internal/external frame of reference model (I/E model), Marsh (1986) described the juxtaposition of social and dimensional comparisons and their impact on an individual's domain-specific academic self-concepts. The I/E model is quite popular in research on self-concept development. In their meta-analysis on 69 studies, Möller, Pohlmann, Köller, and Marsh (2009) found strong support for the central assumptions of the I/E model. For example, the findings indicated that the effects of social and dimensional comparisons are not restricted to specific age groups, gender groups, or countries. The potential moderators investigated in the meta-analysis largely failed to explain the variance in effects between studies, which existed regarding the cross-domain effects in particular. It is interesting to note that there is little research explicitly investigating moderation to explain the variance in effects in the I/E model at the student level. The purpose of this study was to investigate the conditions under which the achievements of others and one's own achievements in different domains contribute to the formation of domain-specific academic self-concepts. To that end, we introduced into the I/E model students' perceptions of their teachers' competence to assess achievement as a potential moderator of the relationship between achievement and self-concept. The idea is that students who perceive the reliability of their teacher-assigned grades as low are less inclined to use social comparison information, which is based on these grades but, due to the need of other information sources, more strongly use dimensional comparison information. In the following sections, we first describe the social and dimensional comparison processes posited by the I/E model and second describe teachers' competence to diagnose achievement.

The Internal/External Frame of Reference Model

The I/E model describes the joint operation of social and dimensional comparisons of students' achievements forming the students' self-concepts (Marsh, 1986). Students conduct social comparisons by comparing their achievement with the achievement of their classmates. If students' verbal achievement is higher than that of classmates, their verbal self-concept is likely to be higher as well. Students conduct dimensional comparisons by comparing their achievement in a given subject with their own achievement in other subjects. For example, if students' verbal achievement is lower than their math achievement, verbal self-concept will suffer and their math self-concept will benefit.

In a meta-analysis of 69 studies on the I/E model, Möller et al. (2009) determined that the average correlation between math and verbal achievement is both positive and strong ($r = .67$) and paradoxically much higher than the average correlation between math and verbal self-concept ($r = .10$). As a result of social comparisons, the two within-domain paths relating math achievement to math self-concept ($\beta = .61$) and verbal achievement to verbal self-concept ($\beta = .49$) were strong and positive as well. The results were the opposite for the two cross-domain paths from verbal achievement to math self-concept ($\beta = -.27$) and from mathematics achievement to verbal self-concept ($\beta = -.21$) as a result of dimensional comparisons. The juxtaposition of dimensional and social comparisons explains the paradoxical finding that self-concepts in different domains are only weakly correlated despite strongly correlated achievements.

Additional findings from longitudinal, experimental, and diary studies have further substantiated these findings (Marsh et al., 2014; Möller & Husemann, 2006; Möller & Köller, 2001; Pohlmann & Möller, 2009; Strickhouser & Zell, 2015). However, moderators of the effects of social and dimensional comparisons have seldom been investigated. Insights into the conditions under which the comparison processes in the I/E model are particularly valid are therefore very limited.

The meta-analysis of Möller et al. (2009) provides some evidence on moderation of effects in the I/E model. Age, gender, country of residence, and type of achievement and self-belief measures were considered as potential moderators in addition to sample size and year of publication of the included studies. Regarding the four path coefficients that are central to the evaluation of the I/E model, multigroup analyses revealed significant moderation of the positive within-domain paths by (a) the type of achievement measure (grades assigned by teachers vs. standardized tests) and (b) the type of self-belief measure (academic self-concept vs. self-efficacy). The only significant moderation regarding the negative cross-domain paths involved the path from math achievement to verbal self-concept, and this was due to the sample size of the data sets as a methodologically relevant moderator.

Little research has explicitly investigated moderators of paths in the I/E model. Skaalvik and Rankin (1992) found significant negative cross-domain paths among students who perceived their levels of math and verbal achievement to be substantially different; in a sample of students who perceived their math and verbal achievement to be relatively equal, these negative paths failed to reach significance. The group differences suggest that dimensional comparisons are moderated by students' perceived relationship between achievements in the contrasting domains. In a similar vein, students' theories about math and verbal abilities in general, that is, whether they believe that both are negatively interdependent or rather independent of each other, moderated dimensional comparisons (more consistently than did social comparisons; Möller, Pohlmann, Streblow, & Kaufmann, 2002; Möller, Streblow, & Pohlmann, 2006): The cross-domain paths in the I/E model were more pronounced when students believed that verbal and math skills were rather distinct. Using an experimental research design, Helm, Müller-Kalthoff, Nagy, and Möller (2016) demonstrated the causal effect of perceived subject similarity on domain-specific self-concepts: Lower perceived subject similarity led to stronger dimensional comparison effects than did higher perceived subject similarity. Steinmayr and Spinath (2015) introduced students' intelligence as a moderator of effects in the I/E model. A multigroup analysis with students assigned to two different groups by intelligence, using the median-split method, showed moderation regarding the cross-domain path from math achievement to verbal self-concept, which was stronger for highly compared with lowly intelligent students. The authors concluded that the importance students assign to their ability in math versus the verbal domain is related to their intelligence.

To sum up, knowledge concerning moderators of the social and dimensional comparison effects is sparse. On the one hand, students' general perceptions of similarity or dissimilarity of math and verbal school subjects may play a role. On the other hand, students evaluate their own ability on the basis of their teachers' feedback, which is why perceived teacher characteristics such as their competence to diagnose achievement may also play a role. It is an open question what happens to the social and dimensional comparison effects when students doubt their teachers' feedback.

Teachers' Diagnostic Competence Concerning Achievement

Teachers' judgments about students' achievements are a main source of information for students as they evaluate their academic achievement and draw conclusions about their academic abilities. Research on the I/E Model and beyond has underscored the role of teachers' assessments of achievement in students' self-related cognitions (Möller et al., 2009; Zimmermann, Schütte, Taskinen, & Köller, 2013). What are the consequences for students of teachers' accuracy in judging their achievement? Urhahne, Chao, Florineth, Luttenberger, and Paechter (2011), for example, showed that misjudgment of students' mathematics achievement had consequences for the students' academic self-concept, expectations of success, and test anxiety. The social and dimensional comparison processes described in the I/E model may depend on the extent to which the feedback provided by grades from a teacher is perceived as a reliable and credible basis on which students can evaluate their own abilities.

Social comparison theory (Festinger, 1954) posits that people strive to evaluate their abilities. The motivations for social comparisons may differ (Festinger, 1954). People may be motivated to (a) evaluate their strengths and weaknesses accurately to gain an understanding of their behavioral effectiveness (Anderson & Lux, 2004; Mead, 1934; Ruble, Boggiano, Feldman, & Loebl, 1980), (b) maintain their self-view or (c) enhance their self-image when comparing themselves with others (Tesser, 1988), or (d) improve their abilities by learning from more accomplished students (Dijkstra, Kuyper, van der Werf, Buunk, & van der Zee, 2008). With social comparison theory setting the stage for the recently introduced dimensional comparison theory, Möller and Marsh (2013) postulated that these motivations also apply to dimensional comparisons.

Whatever the initial motivation for self-evaluation, individuals need access to accurate and valid information if social and dimensional comparisons are to be meaningful. Interindividual differences in perceptions of a teacher's competence to assess student achievement should therefore be associated with the use of social comparisons based on the grades teachers assign to students. For example, students who perceive that their math teacher is not really capable of diagnosing their math achievement should be less inclined to use the grades that the teacher gives as a basis for evaluating their own skills in mathematics and be more inclined to use other sources to evaluate their ability. For example, students might turn to a comparison of their math and verbal achievement.

By and large, teachers' assessments of student achievement are fairly accurate. A meta-analysis of 73 effect sizes on teachers' judgment accuracy found an overall mean effect size of $z_r = .63$ and a range between −.03 and 1.18 (Südkamp, Kaiser, & Möller, 2012). Obviously, there is strong interindividual variance, as well as room for improvement in teachers' judgment accuracy (see also Helmke & Schrader, 1987; Hoge & Coladarci, 1989). The interindividual variance in teachers' competence to diagnose achievement may be reflected in the teachers' diagnostic competence as perceived by the students or may be even greater given the subjectivity of their individual perceptions. How students subjectively perceive their teachers' competence in assessing achievement is a crucial factor in determining whether students will take into account the feedback they receive from their teachers. This study therefore focuses on students' perceptions of teachers' competence in assessing achievement, which reflects students' beliefs about the degree to which their subject teachers are capable of assessing their achievements. The question is whether differences in students' perceptions in this regard have consequences for their self-concepts.

To sum up, although individuals may be motivated by different reasons to evaluate their abilities by comparing achievements between students or domains, they will in any case base their judgments on salient information they consider to be sufficiently reliable. Our study thus focuses on students' perceptions of teachers' diagnostic competence as a key moderator of those motivational processes. If students perceive a teacher to be competent to assess their performance in a given domain, one would expect the grades given by that teacher to contribute to the formation of their self-concept within that domain. Conversely, a lower level of perceived teacher competence should lead students to rely more on other sources in forming their self-concept. In particular, one would expect the achievement in another salient subject domain to become more predictive of self-concept.

Specifically, we hypothesized that a lower level of perceived diagnostic competence concerning math (verbal) achievement (a) weakens the effect of the math (verbal) grade on math (verbal) self-concept (social comparisons) and (b) strengthens the effect of the verbal (math) grade on math (verbal) self-concept in the I/E model (dimensional comparisons).

Method

Samples and Procedure

Participants of the main study (Sample 1) were $N = 1{,}045$ seventh graders ($M_{\text{age}} = 12.7$ years, $SD = .7$; 50% female) in 55 classes from 10 urban schools or school centers covering the whole range of academic and vocational tracks of the segregated school system in Germany. They were the main cohort of the larger multicohort longitudinal project EIKA (Development and Implementation of a School-to-Work Transition Concept for Schools Serving Disadvantaged Communities). This project originally focused on the conditions and determinants of competencies among students from disadvantaged communities, which were drawn to represent an urban area that performed poorly in large-scale international student assessments. To make use of the possibility to replicate the study's main findings in a different sample (Sample 2), we used an independent sample of seventh graders from further cohorts of the same project ($N = 1{,}966$ from 87 classes; $M_{\text{age}} = 12.8$ years, $SD = .7$; 48% female).

The present article is based on self-reported data from seventh graders and on their report card grades from the end of Grade 6. The students' families were characterized by low to moderate socioeconomic status, based on the Highest International Socio-Economic Index of Occupational Status [HISEI] of the mother and father as reported by the parents ($M = 39$, $SD = 14$ in Sample 1 and $M = 46$, $SD = 14$ in Sample 2; on a scale from 16 to 90, with higher scores indicating higher socioeconomic status; Ganzeboom & Treiman, 1996; $M_{\text{HISEI}} = 49$, $SD = 16$, for countries in the Organisation for Economic Co-operation and Development; Klieme et al., 2010). About 50% of the samples comprised students with an immigration background (defined by whether at least one of the parents or the child was not born in Germany), according to the country of birth reported by the parents (or by the children if parents' reports were not available).

Measures

Mathematics achievement (MACH) and *verbal achievement* (VACH) were assessed by means of students' school grades in mathematics and German. That information was collected from students' report cards at the end of Grade 6. The German grading system includes grades from *outstanding* (1) to *fail* (6). To facilitate the interpretation of results, we reverse-coded school grades so that higher scores reflected better outcomes.

Students' ratings were collected in the middle of Grade 7. Domain-specific *mathematics self-concept* (MSC) and *verbal self-concept* (VSC) were defined in terms of students' general belief that they were doing well or poorly in that domain. Mathematics and verbal self-concepts were each measured by four items selected from previously developed scales (Jerusalem, 1984; Jopt, 1978). Sample items include "I am not particularly good in mathematics" and "Nobody's perfect. I am just not talented in German." Responses were rated on a 4-point scale ranging from 1 (*agree*) to 4 (*disagree*), with higher scores indicating a more favorable self-concept. Reliabilities were good; Sample 1: Cronbach's α/McDonald's ω = .82/.85 (MSC) and .82/.86 (VSC): Sample 2: Cronbach's α/McDonald's ω = .86/.86 (MSC) and .86/.85 (VSC). The measurement model for the self-concepts fit the data well; Sample 1: $\chi^2(15) = 82.23$, comparative fit index (CFI) = .97, Tucker–Lewis index (TLI) = .95, root-mean-square error of approximation (RMSEA) = .06, standardized root-mean-square residual (SRMR) = .04; Sample 2: $\chi^2(15) = 51.93$, CFI = .99, TLI = .99, RMSEA = .04, SRMR = .02.

Students' perceptions of their teachers' diagnostic competence concerning mathematics achievement (PDC-MACH) and *verbal achievement* (PDC-VACH) were assessed using a five-item scale developed in the course of the BIJU study (Learning Processes, Educational Careers, and Psychosocial Development in Adolescence and Young Adulthood; Baumert, Gruehn, Heyn, Köller, & Schnabel, 1997; partially based on Fend & Specht, 1986). The items were "Our math [German] teacher . . ." ". . . knows exactly how each of us performs," ". . . notices immediately if one does not understand something," ". . . knows immediately which tasks are difficult for us," ". . . knows immediately what one has not understood," and ". . . notices immediately when a student cannot keep up in class." Responses were rated on a 4-point scale ranging from 1 (*agree*) to 4 (*disagree*). To facilitate the interpretation of results, we reverse-coded the items so that higher scores represented greater perceived competence. Reliabilities were good; Sample 1: Cronbach's α/McDonald's ω = .85/.85 (PDC-MACH) and .92/.91 (PDC-VACH); Sample 2: Cronbach's α/McDonald's ω = .88/.88 (PDC-MACH) and .92/.91 (PDC-VACH). The measurement model for the perceived diagnostic competence scales fit the data well; Sample 1: $\chi^2(29) = 108.87$, CFI = .98, TLI = .97, RMSEA = .05, SRMR = .03; Sample 2: $\chi^2(29) = 124.90$, CFI = .99, TLI = .98, RMSEA = .04, SRMR = .02. Furthermore, we conducted a series of multigroup confirmatory factor models of ascending restrictiveness in each of the two samples to test for the scales' measurement invariance across different school tracks (e.g., Meredith & Teresi, 2006; Steinmetz, Schmidt, Tina-Booh, Wieczorek, & Schwartz, 2009). According to the recommended criteria of differences in CFI < .01 and RMSEA < .015 (Chen, 2007; Cheung & Rensvold, 2002), the model assuming metric measurement invariance fit the data equally as well as did the least restrictive baseline model of configural measurement invariance (Sample 1: ΔCFI = .001, ΔRMSEA = .001; Sample 2: ΔCFI = .000 and ΔRMSEA = .002), and the more restrictive model of scalar measurement invariance fit equally as well as did the model assuming metric measurement invariance (Sample 1: ΔCFI = .003, ΔRMSEA = .000; Sample 2: ΔCFI = .002, ΔRMSEA = .001). The well-fitting measurement models, which functioned equally well in different school tracks, prove the scales' factorial validity and support the comparability regarding their meaning across academic and nonacademic track schools. Moreover, although small to moderate associations are plausible, students should not merely state that a teacher is not competent to diagnose achievement just because they feel unjustly treated or simply do not like the teacher. Correlations in the sample of seventh graders from the BIJU study show that the latent factor of perceived teachers' diagnostic competence concerning math achievement is positively related with but different from liking the teacher ("I like our math teacher very much", Baumert et al., 1997; $n = 5{,}420$; $r = .45$, $p \leq .001$) and feelings of justice concerning the given grades ("Our math teacher assigns grades fairly", Baumert et al., 1997; $n = 5{,}821$; $r = .28$, $p \leq .001$), which supports the validity of the construct. In addition, factor analyses from prior research using the perceived teachers' diagnostic competence scales regarding math, physics, and biology from the BIJU study support that students differentiate between the different subjects when judging their teachers' diagnostic competences rather than that their judgments are a result of individual perceptual biases (Gruehn, 2000).

Statistical Analyses

Structural equation modeling using robust maximum likelihood estimation was conducted with Mplus 7.0 (Muthén & Muthén, 2012). All multi-item constructs (i.e., MSC, VSC, PDC-MACH, PDC-VACH) were specified as latent factors. For items of the same constructs across subjects, wording was identical except for the name of the subject. Thus, correlated uniqueness between indicators with the same wording was incorporated (Marsh, 1990). Standard errors were adjusted for the nested structure of students within classes using the Mplus option <type is complex> (Hox, 2002). On average, 29%

of the data were missing from the model variables in Sample 1 and 25% in Sample 2. Missing values in each sample were addressed by multiple imputation (m = 10 data sets) based on the model's variables plus auxiliary variables (e.g., sex, immigration background, school track, general cognitive abilities, interest in math and German, self-esteem, interaction terms) in Mplus (Graham, 2009).

Moderation of paths in the I/E model was analyzed by including latent interaction variables between achievement and perceived diagnostic competence using the unconstrained approach (Marsh, Wen, & Hau, 2004). This approach does not require constraints on the factor loadings of the product indicators. As Marsh et al. (2004) have shown, moreover, the unconstrained approach is preferable to constrained approaches when variables are nonnormally distributed. After centering the indicators, we formed cross products by multiplying indicators from the variables involved in a specific interaction. These products were then used as indicators of the latent product variables. Consistent with the unconstrained approach of Marsh et al. (2004), the loadings of the product indicators were unconstrained. The errors of the product indicators that had a common component covaried. Means of the manifest first-order predictors (achievements) were zero because they had been centered, and means of the latent first-order predictors (perceived diagnostic competence) were fixed to zero; means of the latent product variables were equivalent to the covariance between the first-order variables (Steinmetz, Davidov, & Schmidt, 2011). The latent product variables were entered into the model along with the first-order predictors. The latent and manifest exogenous variables as well as the latent dependent variables were allowed to correlate.

Table 1
Descriptives and Correlations for the Study Variables for Sample 1

Variable	*M*	*SD*	1	2	3	4	5
1. MACH^a	3.81	.97	—				
2. VACH^a	3.87	.85	.53***	—			
3. MSC	2.83	.87	.48***	.10	—		
4. VSC	2.86	.81	.10*	.35***	.10	—	
5. PDC-MACH	3.03	.66	.01	−.05	.11*	−.05	—
6. PDC-VACH	3.07	.74	−.08	.04	−.01	.02	.44***

Note. Means are manifest; correlations are between latent variables (except achievements, which are manifest school grades). Possible scale range is from 1 to 4, unless otherwise indicated. MACH = math achievement; VACH = verbal achievement; MSC = math self-concept; VSC = verbal self-concept; PDC-MACH = perceived teachers' diagnostic competence concerning math achievement; PDC-VACH = perceived teachers' diagnostic competence concerning verbal achievement.
^{a} Possible scale range from 1 to 6.
* $p \leq .05$. *** $p \leq .001$.

Results

Sample 1

Descriptive statistics and correlations between the study variables are shown in Table 1. Achievement in mathematics and German correlated positively, as expected, as did achievement and self-concept within each domain. Students' perceptions of teachers' diagnostic competence in mathematics correlated modestly with students' mathematics self-concept, whereas there was no such correlation in the verbal domain. Teachers' perceived diagnostic competence in mathematics and in the verbal domain correlated moderately with each other; each was uncorrelated with students' achievement.

The hypothesized moderated I/E model (see Figure 1) fit the data well, $\chi^2(684)$ = 1,285.54, CFI = .96, TLI = .96, RMSEA = .03, SRMR = .03. Predictions of the I/E model were fully supported. Paths from achievement to the corresponding self-concept within a domain were positive, indicating social comparison effects; cross-domain paths were negative, indicating dimensional comparison effects. More important, regarding moderation, two of the four paths from the latent interaction variables were significant. The pattern of results shows that both cross-domain paths were significantly moderated, whereas the within-domain paths were not.

Students' perceptions of their teachers' diagnostic competence for verbal achievement moderated the cross path from mathematics achievement to verbal self-concept. To describe the meaning of this moderation effect, Figure 2 shows the simple slopes for the regression of verbal self-concept on mathematics achievement within the I/E model at different levels of the moderator PDC-VACH: Math achievement contributed most strongly to verbal self-concept when teachers' diagnostic competence for verbal achievement was perceived to be low. The cross path from mathematics achievement to verbal self-concept was smaller the greater a teacher's competence for assessing verbal achievement was perceived to be.

Similarly, perceptions of teachers' diagnostic competence concerning math achievement moderated the cross path from verbal achievement to math self-concept. To describe the meaning of this moderation effect, Figure 3 depicts the simple slopes for the regression of math self-concept on verbal achievement within the I/E model at different levels of the moderator PDC-MACH: Verbal achievement contributed most strongly to mathematics self-concept when teachers' diagnostic competence for mathematics achievement was perceived to be low. The cross path from verbal achievement to mathematics self-concept fell in size with increases in teachers' perceived diagnostic competence for mathematics achievement. Thus, verbal patterns parallel math patterns, with one exception: The differences in the strength of the cross-domain path from verbal achievement were most pronounced for better verbal grades. That is, among students with good verbal grades, those who perceived their teacher's diagnostic competence in math to be low tended to have a worse math self-concept. Contrast effects appeared to be robust and independent of teachers' perceived diagnostic competence for those with comparatively worse verbal grades.

Sample 2

Descriptive statistics and correlations between the study variables are shown in Table 2. Achievement in mathematics and German correlated positively, as expected, as did achievement and self-concept within each domain. Students' perceptions of teachers' diagnostic competences were not significantly related to their self-concepts. Teachers' perceived diagnostic competence in mathematics and in the verbal domain correlated moderately with each other. Each were slightly negatively correlated with students' achievements in the cross-domain subjects; perceived diagnostic competence in the verbal domain correlated slightly negatively

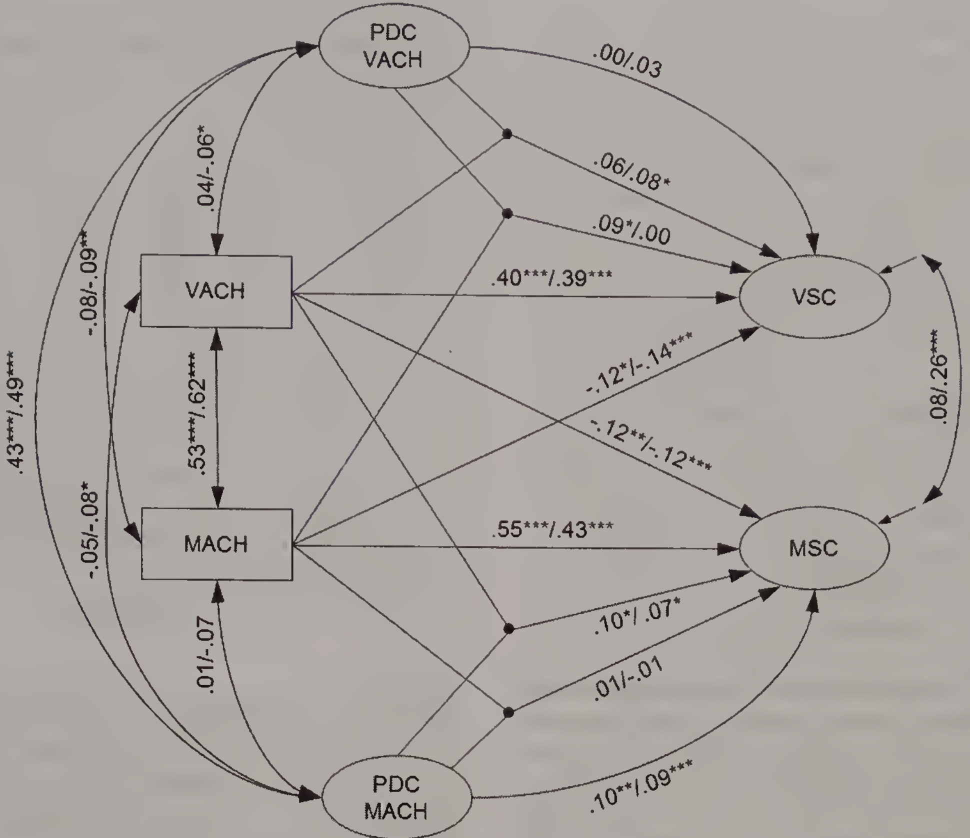


Figure 1. Results for the internal/external frame of reference model, including the hypothesized moderating effects of perceived teachers' domain-specific diagnostic competence for Sample 1 (before the slashes) and Sample 2 (after the slashes). MACH (VACH) = math (verbal) achievement; MSC (VSC) = math (verbal) self-concept; PDC-MACH (PDC-VACH) = perceived teachers' diagnostic competence concerning math (verbal) achievement. Standardized path coefficients. The latent interaction variables are represented by the paths of two interacting variables converging at a dot. * $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

with verbal achievement, whereas the equivalent correlation in the domain of mathematics was not significant.

The hypothesized moderated I/E model (see Figure 1) fit the data well, $\chi^2(684) = 1{,}314.62$, CFI = .98, TLI = .98, RMSEA = .02, SRMR = .02. Predictions of the I/E model were fully supported. Regarding moderation, two of the four paths from the latent interaction variables were significant. First, moderation of the within-domain path from verbal achievement to verbal self-concept reached significance (which was in the same direction but did not reach significance in Sample 1). Second, the moderation of the cross-domain effect on math self-concept was replicated.

Students' perceptions of their teachers' diagnostic competence for verbal achievement moderated the within path from verbal achievement to verbal self-concept. To describe the meaning of this moderation effect, Figure 4 shows the simple slopes for the regression of verbal self-concept on verbal achievement within the I/E model at different levels of the moderator PDC-VACH: Verbal achievement contributed most strongly to verbal self-concept when teachers' diagnostic competence for verbal achievement was perceived to be high. The within path from verbal achievement to verbal self-concept was smaller the lower a teacher's competence for assessing verbal achievement was perceived to be.

Students' perceptions of teachers' diagnostic competence concerning math achievement moderated the cross path from verbal achievement to math self-concept. Figure 5 depicts the simple slopes for the regression of mathematics self-concept on verbal achievement within the I/E model at different levels of the moderator PDC-MACH, and shows virtually the same pattern as Figure 3 for Sample 1: Verbal achievement contributed stronger to mathematics self-concept the lower the perceived teachers' diagnostic competence was for mathematics achievement. Again, the differences were most pronounced for better verbal grades: Among students with good verbal grades, those who perceived their teacher's diagnostic competence in math to be low tended to have a worse math self-concept. Contrast effects appeared to be rather independent of teachers' perceived diagnostic competence for those with comparatively worse verbal grades.

Finally, the path from perceived teacher's diagnostic competence for math achievement to math self-concept, which we did not hypothesize a priori but which was significant in Sample 1, was replicated in Sample 2.

Discussion

Ample research has documented the replicability of effects in the I/E model (Möller et al., 2009). The aim of this study was to obtain a more comprehensive picture of conditions under which students' achievements contribute to the formation of academic self-concepts. To that end, we investigated moderation in the I/E model of paths within domains, which are posited to result from social comparison, and of paths across domains, which are posited to result from dimensional comparison (Marsh, 1986). In accor-

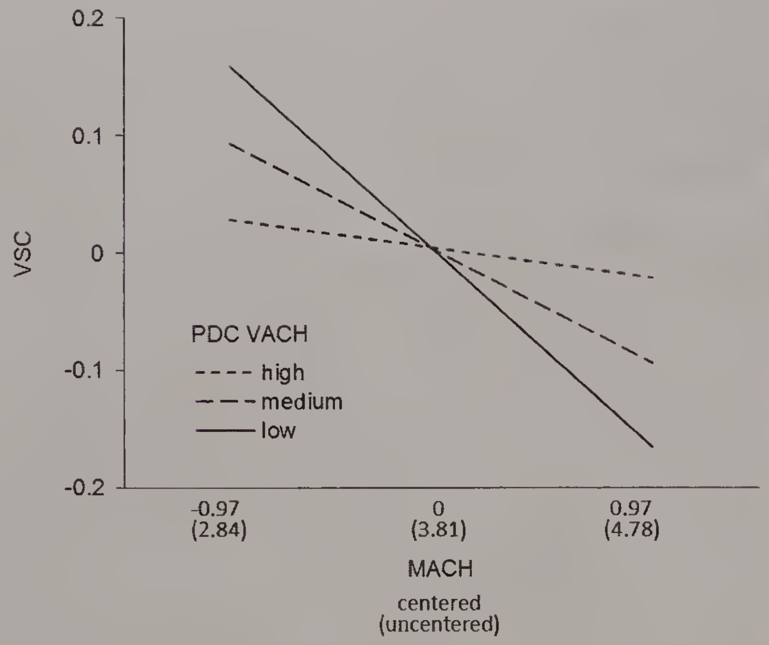


Figure 2. Moderation of the cross-domain effect on verbal self-concept (VSC) for Sample 1. Simple slopes for the influence of math achievement (MACH) on VSC at different levels of perceived teachers' diagnostic competence concerning verbal achievement (PDC-VACH) based on centered data (high = *M* + 1 *SD*, medium = *M*, low = *M* – 1 *SD;* Cohen, Cohen, West, & Aiken, 2003). Units of the *y*-axis are in terms of standard deviation.

Table 2
Descriptives and Correlations for the Study Variables for Sample 2

Variable	*M*	*SD*	1	2	3	4	5
1. MACH[a]	3.65	1.07	—				
2. VACH[a]	3.71	.95	.62***	—			
3. MSC	2.78	.88	.31***	.10***	—		
4. VSC	2.82	.81	.08**	.29***	.24***	—	
5. PDC-MACH	2.96	.72	−.07	−.08*	.06	−.05	—
6. PDC-VACH	3.00	.76	−.09**	−.06*	−.01	.01	.50***

Note. Means are manifest; correlations are between latent variables (except achievements, which are manifest school grades). Possible scale range is from 1 to 4, unless otherwise indicated. MACH = math achievement; VACH = verbal achievement; MSC = math self-concept; VSC = verbal self-concept; PDC-MACH = perceived teachers' diagnostic competence concerning math achievement; PDC-VACH = perceived teachers' diagnostic competence concerning verbal achievement.
[a] Possible scale range from 1 to 6.
$^{*}p \leq .05$. $^{**}p \leq .01$. $^{***}p \leq .001$.

dance with the theories of social and dimensional comparison (Festinger, 1954; Möller & Marsh, 2013), we examined as moderators students' perceptions of their teachers' competences in diagnosing their achievements. Drawing on two large samples, we consistently found that the cross-domain effect on math self-concept was moderated: The verbal grade related stronger to math self-concept the lower students perceived their teachers' competence in diagnosing math achievement. Furthermore, the results of our studies reasonably converge in that the within-domain effect of the verbal grade on verbal self-concept is stronger with higher perceived diagnostic competence for verbal achievement.

Moderation of Cross-Domain Paths in the I/E Model

The current study has shown that dimensional comparison effects, in particular the contrast effect on math self-concept when comparing verbal with math achievement, decreases when students believe that their math teacher is highly competent in diagnosing

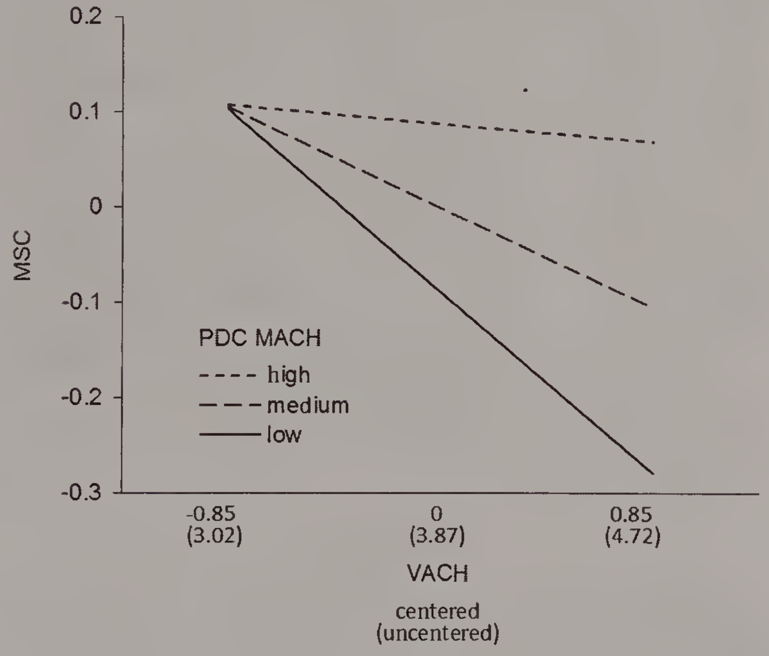


Figure 3. Moderation of the cross-domain effect on math self-concept (MSC) for Sample 1. Simple slopes for the influence of verbal achievement (VACH) on MSC at different levels of perceived teachers' diagnostic competence concerning math achievement (PDC-MACH) based on centered data (high = *M* + 1 *SD*, medium = *M*, low = *M* – 1 *SD;* Cohen et al., 2003). Units of the *y*-axis are in terms of standard deviation.

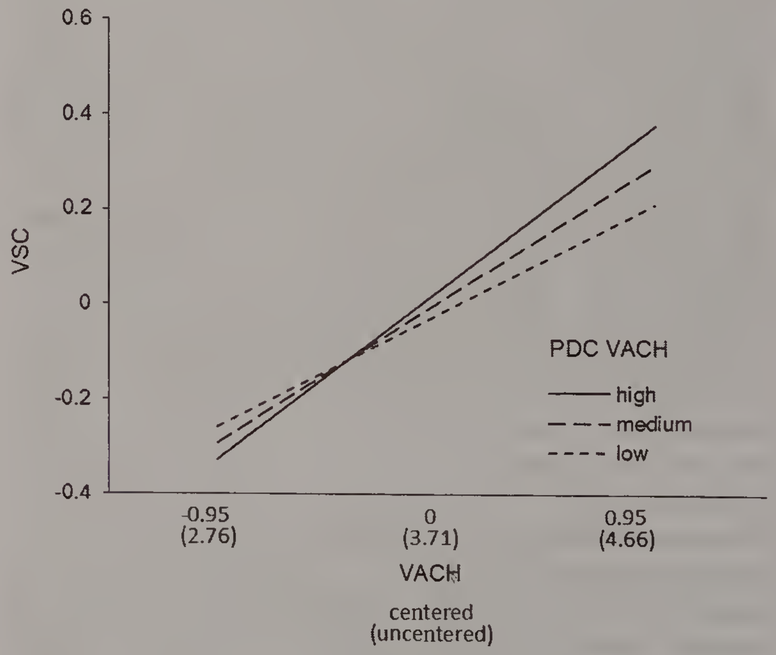


Figure 4. Moderation of the within-domain effect on verbal self-concept (VSC) for Sample 2. Simple slopes for the influence of verbal achievement (VACH) on VSC at different levels of perceived teachers' diagnostic competence concerning verbal achievement (PDC-VACH) based on centered data (high = *M* + 1 *SD*, medium = *M*, low = *M* – 1 *SD;* Cohen et al., 2003). Units of the *y*-axis are in terms of standard deviation.

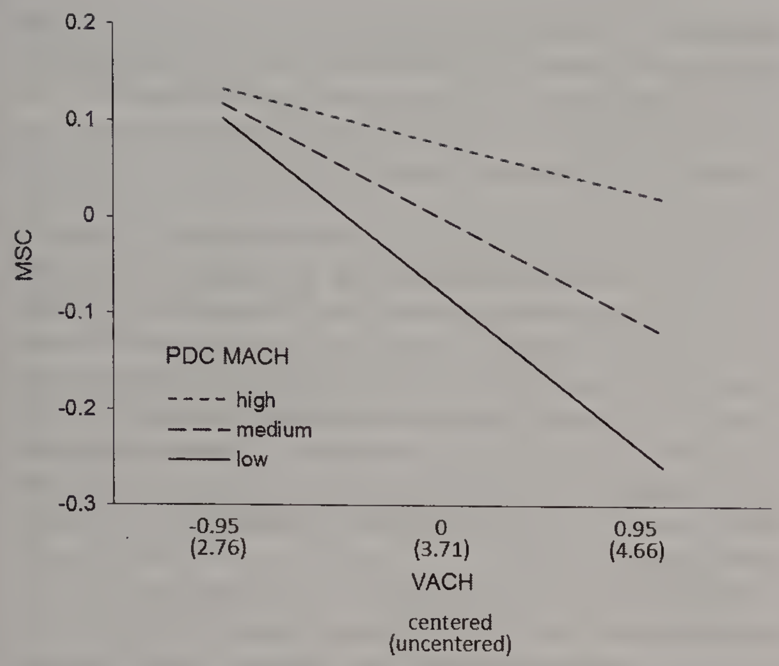


Figure 5. Moderation of the cross-domain effect on math self-concept (MSC) for Sample 2. Simple slopes for the influence of verbal achievement (VACH) on MSC at different levels of perceived teachers' diagnostic competence concerning math achievement (PDC-MACH) based on centered data (high = $M + 1$ *SD*, medium = *M*, low = $M - 1$ *SD;* Cohen et al., 2003). Units of the *y*-axis are in terms of standard deviation.

achievement. The finding is as expected and suggests that verbal achievement is less important for the math self-concept when the teachers' diagnostic competence in math is believed to be high. Conversely, students' math self-concept tends to be particularly influenced by verbal achievement when students believe that their math teachers' diagnostic competence is low. Therefore, the effects of dimensional comparisons related to grades on domain-specific self-concept seem to be (at least in part) a consequence of a low level of trust in teachers' diagnostic competence. Accordingly, math teachers' perceived diagnostic competence is taken into account when students evaluate their abilities in math based on dimensional comparisons of achievements.

In accordance with the predominantly self-serving nature of dimensional comparisons, students may exaggerate their abilities in a given domain by comparing them with worse grades in a contrasting domain (cf. Pohlmann & Möller, 2009). Our studies show that the math self-concept of those with worse verbal grades generally benefits from the contrast effect rather independently of their math teacher's perceived diagnostic competence. We have consistently found that students with better verbal grades suffer from negative contrasting effects on their math self-concept, particularly when they perceive their math teacher's diagnostic competence to be low: Students with good verbal grades tend to have a worse opinion of their math skills the less they trust their math teacher's competence in diagnosing achievement. Thus, students might give relatively more weight to other sources such as contrasting subject domains by means of dimensional comparison when they perceive their teacher's competence to assess achievement in the relevant domain to be low.

Moderation of Within-Domain Paths in the I/E Model

The within-domain effects, which are assumed to result from social comparison processes, were not consistently significantly moderated. The results of our studies tentatively imply that the perceived competence of teachers to diagnose achievement contributes to the degree that students incorporate their grades into their self-concepts in the verbal domain. To us, it is somewhat surprising that the moderation of the within-domain paths yielded only sparse results, which were even less pronounced over our studies than was the moderation of the cross-domain paths. We postulated that students' perceptions of their teachers' competence to diagnose achievement were a prerequisite for social comparisons based on grades given by teachers.

However, some reasons may explain the mostly nonsignificant within-domain paths in the I/E model. Social comparisons may be initiated immediately and almost automatically when achievement feedback is available for oneself and relevant others. For example, school grades are salient in the classroom when exam results are returned in a subject. Explicitly or implicitly, students surely attach great importance to their teachers' assessments of academic achievement as reflected in school grades. These teacher assessments are an important currency in school when it comes to decisions regarding students' further educational careers within and beyond school (Harlen, 2005). The availability of information about school grades in a particular subject, along with the significance of grades in the corresponding subject, may result in relatively robust social comparisons even in the presence of doubts about their accuracy.

The less significant results regarding moderation of the within-domain paths are also more or less in line with findings from previous studies that have investigated moderation of within- and cross-domain paths in the I/E model at the student level. Skaalvik and Rankin (1992) did not report any moderation of the within-domain paths by students' perceived similarity of their math and verbal achievement. Nor was there evidence of a moderation of the social comparison processes within the I/E model by students' intelligence (Steinmayr & Spinath, 2015). Finally, research by Möller and colleagues yielded only inconsistent evidence for a moderation of social comparison processes by students' beliefs about the negative interdependency of math and verbal abilities (Möller et al., 2002, 2006).

To conclude, the perceived validity of school grades may encourage students to rely on this achievement feedback or, if perceived to be low, on achievement in other domains when evaluating themselves. Considering the bigger picture, it has to be kept in mind that these are only some aspects of the information that may flow into the formation of students' self-concepts. An exciting challenge for self-concept research is the fact that the various frames of reference and other sources of information for self-beliefs are often in place simultaneously. For example, the total effect of an exam on domain-specific self-concept will depend on how a student weighs all these pieces of comparative information. Not surprisingly, the comparison processes (i.e., social, dimensional, temporal comparisons, and comparison with a criterion) and their interdependence (see, e.g., Wakslak, Nussbaum, Liberman, & Trope, 2008; Wilson & Ross, 2000; Zell & Alicke, 2009) as well as other sources of information for self-beliefs like the average class or school ability (Chmielewski, Dumont, & Trautwein, 2013; Marsh, 1987), the reputation of the school (Marsh, Kong, & Hau, 2000), socializers' feedback (Pohlmann, Möller, & Streblow, 2004), expectancy effects (Jussim & Harber, 2005), and gender stereotypes (Watt & Eccles, 2008) continue to be the subject of empirical investigation and conceptual refinement.

Teachers' Perceived Diagnostic Competence Concerning Achievement

When it comes to students' perceptions of their teachers' competence to assess achievement, there are clear overlaps between the mathematics and verbal domains. This is most likely due to personal characteristics of the students. Accordingly, the correlations between students' perceptions of the diagnostic competence of teachers from different subjects may be the result of a more general belief in teachers' diagnostic competence, independent of the specific teacher or domain. However, individuals' subjective experience—based on the lens through which they see the world—is crucial for their self-evaluation.

Note also that the perceived teachers' diagnostic competences and the grades provided by the teachers in the same domain were mostly uncorrelated (with one exception of a small and even negative correlation in the verbal domain in Sample 2). This suggests that students rate their teachers' diagnostic competence in a given domain quite independently of their own achievement in that domain. So although students' assessments are probably influenced by their personal characteristics and beliefs and are thus by no means objective (which is neither to be expected nor necessary to be powerful), they are not biased in the sense that students with better school grades rate their teachers' diagnostic competence better.

An additional finding from both studies is that when students perceive their math teachers to be highly competent to assess their math achievement, this is generally associated with a more favorable math self-concept. An adequate level of diagnostic competence concerning their students' achievements is regarded as necessary for teachers to conduct instruction that promotes individual learning progress (Hattie, 2009). This also involves providing helpful individual feedback about performance level and knowledge gaps or to judge students' performance based on an individual frame of reference. Previous research in the domain of mathematics has shown that teachers' use of an individual frame of reference as perceived by students is positively associated with their self-concepts (Lüdtke, Köller, Marsh, & Trautwein, 2005). It is therefore possible that students' math self-concepts are higher, the more they perceive that their math teachers judge achievement highly based on an individual frame of reference, which should be linked to their perception that the math teachers are highly competent in diagnosing math achievement. In contrast, this correlation was consistently nonsignificant in the verbal domain, perhaps because of the high level of heterogeneity of the verbal self-concept (Arens & Jansen, 2016).

Educational Implications

Our results suggest that most students believe that their teachers are reasonably competent to diagnose achievements in their subjects (math and German). Meta-analyses of teachers' judgment accuracy support the students' perceptions (Hoge & Coladarci, 1989; Südkamp et al., 2012). Such objectively and perceived high diagnostic competences of teachers are desirable, because misjudgments may have detrimental consequences for students' self-concepts, emotions, and behavior (e.g., Urhahne et al., 2011; Zimmermann et al., 2013). Supporting the development of an appropriate academic self-concept is in turn an important goal in education, given its implications for interest, academic behavior, and achievement (e.g., Eccles, 1983; Valentine et al., 2004).

The current research shows that the integration of students' verbal achievement into their verbal self-concept is somewhat reduced, whereas the distorting effect of dimensional comparison by incorporating verbal achievement into the math self-concept (Möller & Marsh, 2013) is amplified by the perception that teachers are not well equipped to assess achievement. In contrast, when teachers are viewed as highly competent to diagnose performance, students are more likely to evaluate their abilities more accurately and less likely to exaggerate their strengths or weaknesses.

Teacher training should provide appropriate learning opportunities for prospective and practicing teachers. They should be given information about diagnosing achievement, possible sources of bias, strategies to minimize such biases, and providing appropriate achievement feedback to students.

Limitations and Directions for Future Research

Among the study's strengths are reasonably large sample sizes, with the power to detect interaction effects, and sophisticated statistical analyses using state-of-the-art latent modeling of interactions. However, it is not without certain limitations. The design of the studies is not truly longitudinal, which precludes the interpretation of results in terms of causal effects. The participants in our studies consisted of a comparatively high proportion of students with an immigration background and were on average from families with a low to moderate socioeconomic background. It is therefore an open question whether the results replicate to student samples with a different composition. Coefficients of the dimensional comparison paths in the I/E model were somewhat lower in these studies compared with those found on average in the meta-analysis by Möller et al. (2009). It is uncertain whether these relatively weak weights indicate the presence of moderators only in our particular samples or whether the phenomenon is more generally valid. A further issue, which the study leaves open for future research, is the generalizability of the results to different domains other than the verbal and mathematics domains. Moreover, it still appears puzzling to us that not all of the interactions yielded significant effects. The interaction effects found were not all replicated, and those significant were rather small. Although there is theoretical importance, it is therefore questionable whether these effects are of great importance, and the practical relevance is probably low. All in all, additional research is needed, and we hope to initiate future research to accumulate more substantial knowledge on the issue of conditions under which students' achievements contribute to the formation of academic self-concepts.

The present research serves as one step toward deeper investigations into the moderation of effects in the course of self-concept formation. Regarding further research on the moderation of effects in the I/E model, it would be interesting to examine student perceptions of teacher characteristics other than the competence to diagnose achievement. For example, the relationships between teachers and students may impact on the relationships between grades and self-concepts. Imagine students who dislike their math teacher. They might disassociate themselves from the math grades, leading to a lower within-domain effect and a stronger cross-domain effect compared with those who like their math teacher. Another candidate for moderation of effects in the I/E model could be the belief in fairness of teacher-assigned grades. If grades are perceived as being assigned unjustly in class in a domain, this might lead to lower within-domain effects and stronger cross-domain effects.

Because our additional analyses of data from the BIJU study indicate that there is a certain (small) amount of overlap between perceived teachers' competence to diagnose math achievement with liking of the math teacher and with feelings of justice regarding grade assignment in math, it would also be important to take into account intercorrelations of these variables within one moderation model to examine their unique contributions.

The present research expands the understanding of the conditions under which within- and cross-domain paths in the I/E model are valid. The results of our studies suggest that within- and cross-domain effects (specifically, the contribution of verbal grades to the math self-concept and also somewhat to the verbal self-concept) are contingent upon the students' perceptions of a teacher characteristic, namely, the competence to diagnose achievement in math or in the verbal domain. In addition, this study contributes to research on teachers' judgment accuracy concerning achievement by taking a closer look at how students' perceptions of their teachers' competence in this area relate to students' self-concept.

References

Anderson, J., & Lux, W. (2004). Knowing your own strength: Accurate self-assessment as a requirement for personal autonomy. *Philosophy, Psychiatry, & Psychology, 11,* 279–294. http://dx.doi.org/10.1353/ppp.2005.0003

Arens, A. K., & Jansen, M. (2016). Self-concepts in reading, writing, listening, and speaking: A multidimensional and hierarchical structure and its generalizability across native and foreign languages. *Journal of Educational Psychology, 108,* 646–664. http://dx.doi.org/10.1037/edu0000081

Baumert, J., Gruehn, S., Heyn, S., Köller, O., & Schnabel, K. (1997). *Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU) Dokumentation* (Vol. 1) [Educational trajectories and psychosocial development in adolescence (BIJU) documentation]. Berlin, Germany: Max-Planck-Institut für Bildungsforschung.

Biernat, M., & Eidelman, S. (2007). Standards. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (2nd ed., pp. 308–333). New York, NY: Guilford Press.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14,* 464–504. http://dx.doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255. http://dx.doi.org/10.1207/S15328007SEM0902_5

Chmielewski, A. K., Dumont, H., & Trautwein, U. (2013). Tracking effects depend on tracking type: An international comparison of students' mathematics self-concept. *American Educational Research Journal, 50,* 925–957. http://dx.doi.org/10.3102/0002831213489843

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Deci, E. L., & Ryan, R. M. (Eds.). (2002). *Handbook of self-determination research.* New York, NY: The University of Rochester Press.

Dijkstra, P., Kuyper, H., van der Werf, G., Buunk, A. P., & van der Zee, Y. G. (2008). Social comparison in the classroom: A review. *Review of Educational Research, 78,* 828–879. http://dx.doi.org/10.3102/0034654308321210

Eccles, J. S. (1983). Expectancies, values, and academic behaviors. In J. T. Spencc (Ed.), *Achievement and achievement motives* (pp. 75–146). San Francisco, CA: Freeman.

Eccles, J. (2009). Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. *Educational Psychologist, 44,* 78–89. http://dx.doi.org/10.1080/00461520902832368

Fend, H., & Specht, W. (1986). *Erziehungsumwelten: Bericht aus dem Projekt "Entwicklung im Jugendalter"* [Educational environments: Report from the Development in Youth Age project]. Faculty of Social Sciences, University of Konstanz.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7,* 117–140. http://dx.doi.org/10.1177/001872675400700202

Ganzeboom, H. B. G., & Treiman, D. J. (1996). Internationally comparable measures of occu-pational status for the 1988 international standard classification of occupations. *Social Science Research, 25,* 201–239. http://dx.doi.org/10.1006/ssre.1996.0010

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576. http://dx.doi.org/10.1146/annurev.psych.58.110405.085530

Gruehn, S. (2000). *Unterricht und schulisches Lernen: Schüler als Quellen der Unterrichtsbeschreibung* [Teaching and schooling: Students as sources of instruction description]. Münster, Germany: Waxmann.

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education, 20,* 245–270. http://dx.doi.org/10.1080/02671520500193744

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London, United Kingdom: Routledge.

Helm, F., Müller-Kalthoff, H., Nagy, N., & Möller, J. (2016). Dimensional comparison theory: Perceived subject similarity impacts on students' self-concepts. *AERA Open, 2*(2). doi:10.1177/2332858416650624

Helmke, A., & Schrader, F.-W. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. *Teaching and Teacher Education, 3,* 91–98. http://dx.doi.org/10.1016/0742-051X(87)90010-2

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59,* 297–313. http://dx.doi.org/10.3102/00346543059003297

Hox, J. (2002). *Multilevel analysis techniques and applications.* Mahwah, NJ: Erlbaum.

Jerusalem, M. (1984). *Selbstbezogene Kognitionen in schulischen Bezugsgruppen: Eine Längsschnittstudie* [Self-related cognitions in school reference groups: A longitudinal study]. Berlin, Germany: Free University Press.

Jopt, U. J. (1978). *Selbstkonzept und Ursachenerklärung in der Schule* [Self-concept and attribution at school]. Bochum, Germany: Kamp.

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9,* 131–155. http://dx.doi.org/10.1207/s15327957pspr0902_3

Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M., & Stanat, P. (Eds.). (2010). *PISA 2009: Bilanz nach einem Jahrzehnt* [PISA 2009: A decade in retrospect]. Münster, Germany: Waxmann.

Lüdtke, O., Köller, O., Marsh, H. W., & Trautwein, U. (2005). Teacher frame of reference and the big-fish-little-pond effect. *Contemporary Educational Psychology, 30,* 263–285. http://dx.doi.org/10.1016/j.cedpsych.2004.10.002

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal, 23,* 129–149. http://dx.doi.org/10.3102/00028312023001129

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology, 79,* 280–295. http://dx.doi.org/10.1037/0022-0663.79.3.280

Marsh, H. W. (1990). Influences of internal and external frames of reference on the formation of math and English self-concepts. *Journal of Educational Psychology, 82,* 107–116. http://dx.doi.org/10.1037/0022-0663.82.1.107

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1,* 133–163. http://dx.doi.org/10.1111/j.1745-6916.2006.00010.x

Marsh, H. W., Kong, C.-K., & Hau, K.-T. (2000). Longitudinal multilevel models of the big-fish-little-pond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality and Social Psychology, 78,* 337–349. http://dx.doi.org/10.1037/0022-3514.78.2.337

Marsh, H. W., Kuyper, H., Seaton, M., Parker, P. D., Morin, A. J. S., Möller, J., & Abduljabbar, A. S. (2014). Dimensional comparison theory: An extension of the internal/external frame of reference effect on academic self-concept formation. *Contemporary Educational Psychology, 39,* 326–341. http://dx.doi.org/10.1016/j.cedpsych.2014.08.003

Marsh, H. W., Wen, Z., & Hau, K.-T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods, 9,* 275–300. http://dx.doi.org/10.1037/1082-989X.9.3.275

Mead, G. H. (1934). *Mind, self, and society.* Chicago, IL: University of Chicago Press.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*(11, Suppl. 3), S69–S77. http://dx.doi.org/10.1097/01.mlr.0000245438.73837.89

Möller, J., & Husemann, N. (2006). Internal comparisons in everyday life. *Journal of Educational Psychology, 98,* 342–353. http://dx.doi.org/10.1037/0022-0663.98.2.342

Möller, J., & Köller, O. (2001). Dimensional comparisons: An experimental approach to the internal/external frame of reference model. *Journal of Educational Psychology, 93,* 826–835. http://dx.doi.org/10.1037/0022-0663.93.4.826

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research, 79,* 1129–1167. http://dx.doi.org/10.3102/0034654309337522

Möller, J., Pohlmann, B., Streblow, L., & Kaufmann, J. (2002). Die Spezifität von Begabungsüberzeugungen als Determinante des verbalen und mathematischen Begabungsselbstkonzepts [Specificity of ability beliefs as a determinant of academic self-concepts]. *Zeitschrift für Pädagogische Psychologie, 16,* 87–97. http://dx.doi.org/10.1024//1010-0652.16.2.87

Möller, J., Streblow, L., & Pohlmann, B. (2006). The belief in a negative interdependence of math and verbal abilities as determinant of academic self-concepts. *British Journal of Educational Psychology, 76,* 57–70. http://dx.doi.org/10.1348/000709905X37451

Möller, J., & Marsh, H. W. (2013). Dimensional comparison theory. *Psychological Review, 120,* 544–560. http://dx.doi.org/10.1037/a0032459

Mussweiler, T., Rüter, K., & Epstude, K. (2006). The why, who, and how of social comparison: A social-cognition perspective. In S. Guimond (Ed.), *Social comparison and social psychology: Understanding cognition, intergroup relations, and culture* (pp. 33–54). New York, NY: Cambridge University Press.

Muthén, L. K., & Muthén, B. O. (2012). Mplus version 7.0 [Computer software]. Los Angeles, CA: Author.

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95,* 667–686. http://dx.doi.org/10.1037/0022-0663.95.4.667

Pohlmann, B., & Möller, J. (2009). On the benefit of dimensional comparisons. *Journal of Educational Psychology, 101,* 248–258. http://dx.doi.org/10.1037/a0013151

Pohlmann, B., Möller, J., & Streblow, L. (2004). Zur Fremdeinschätzung von Schülerselbstkonzepten durch Lehrer und Mitschüler [On students' self-concepts inferred by teachers and classmates]. *Zeitschrift für Pädagogische Psychologie, 18,* 157–169.

Ruble, D. N., Boggiano, A. K., Feldman, N. S., & Loebl, J. H. (1980). Developmental analysis of the role of social comparison in self-evaluation. *Developmental Psychology, 16,* 105–115. http://dx.doi.org/10.1037/0012-1649.16.2.105

Skaalvik, E. M., & Rankin, R. J. (1992). Math and verbal achievement and self-concepts: Testing the internal/external frame of reference model. *Journal of Early Adolescence, 12,* 267–279. http://dx.doi.org/10.1177/0272431692012003003

Steinmayr, R., & Spinath, B. (2015). Intelligence as a potential moderator in the internal/external frame of reference model. *Journal for Educational Research Online, 7,* 198–218.

Steinmetz, H., Davidov, E., & Schmidt, P. (2011). Three approaches to estimate latent interaction effects: Intention and perceived behavioral control in the theory of planned behavior. *Methodological Innovations Online, 6,* 95–110. http://dx.doi.org/10.4256/mio.2010.0030

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. *Quality & Quantity: International Journal of Methodology, 43,* 599–616. http://dx.doi.org/10.1007/s11135-007-9143-x

Strickhouser, J. E., & Zell, E. (2015). Self-evaluative effects of dimensional and social comparison. *Journal of Experimental Social Psychology, 59,* 60–66. http://dx.doi.org/10.1016/j.jesp.2015.03.001

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104,* 743–762. http://dx.doi.org/10.1037/a0027627

Suls, J., & Wheeler, L. (Eds.). (2000). *Handbook of social comparison: Theory and research.* http://dx.doi.org/10.1007/978-1-4615-4237-7

Tesser, A. (1988). Toward a self-evaluation maintenance model of social behavior. *Advances in Experimental Social Psychology, 21,* 181–227. http://dx.doi.org/10.1016/S0065-2601(08)60227-0

Urhahne, D., Chao, S.-H., Florineth, M. L., Luttenberger, S., & Paechter, M. (2011). Academic self-concept, learning motivation, and test anxiety of the underestimated student. *British Journal of Educational Psychology, 81,* 161–177. http://dx.doi.org/10.1348/000709910X504500

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist, 39,* 111–133. http://dx.doi.org/10.1207/s15326985ep3902_3

Wakslak, C. J., Nussbaum, S., Liberman, N., & Trope, Y. (2008). Representations of the self in the near and distant future. *Journal of Personality and Social Psychology, 95,* 757–773. http://dx.doi.org/10.1037/a0012939

Watt, H. M. G., & Eccles, J. S. (Eds.) (2008). *Gender and occupational outcomes: Longitudinal assessments of individual, social, and cultural influences.* http://dx.doi.org/10.1037/11706-000

Wilson, A. E., & Ross, M. (2000). The frequency of temporal-self and social comparisons in people's personal appraisals. *Journal of Personality and Social Psychology, 78,* 928–942. http://dx.doi.org/10.1037/0022-3514.78.5.928

Zell, E., & Alicke, M. D. (2009). Self-evaluative effects of temporal and social comparison. *Journal of Experimental Social Psychology, 45,* 223–227. http://dx.doi.org/10.1016/j.jesp.2008.09.007

Zimmermann, F., Schütte, K., Taskinen, P., & Köller, O. (2013). Reciprocal effects between adolescent externalizing problems and measures of achievement. *Journal of Educational Psychology, 105,* 747–761. http://dx.doi.org/10.1037/a0032793

Received April 13, 2016
Revision received January 24, 2017
Accepted January 28, 2017 ■

Journal of Educational Psychology
2018, Vol. 110, No. 1, 58–71

© 2017 American Psychological Association
0022-0663/18/$12.00 http://dx.doi.org/10.1037/edu0000206

Effects of a Strategy-Focused Instructional Program on the Writing Quality of Upper Elementary Students in the Netherlands

Renske Bouwer, Monica Koster, and Huub van den Bergh
Utrecht University

In this study, the authors tested the effects of Tekster [Texter], a comprehensive strategy-focused writing instruction program, using a switching replication design with three measurement occassions. The program was implemented by fourth, fifth, and sixth grade teachers ($N = 76$) in 60 general education classrooms in the Netherlands. Students ($n = 688$) and teachers ($n = 31$) in Group 1 worked with Tekster during the first 8-week period, between the first and second measurement occasion. Students ($n = 732$) and teachers ($n = 45$) in Group 2 implemented Tekster during the second 8-week period, between the second and third measurement occasion. The intervention led to statistically significant improvements in the quality of students' writing. The effect size for the full sample was 0.32 and 0.40 for students who students who completed all 16 Tekster lessons. Gains shown by students in Group 1 were maintained after 8 weeks. Because writing quality was assessed in 3 genres, the findings are generalizable across students, classes, and writing tasks. Taken together, the results of this study demonstrate that a strategy-focused writing instruction program, such as Tekster, can be an effective way to improve upper-elementary students' written language skills.

Educational Impact and Implications Statement

This study shows how Tekster [Texter], a strategy-focused writing instruction program, improves the writing performance of students in Grade 4 to 6. This positive effect was still visible 2 months after the intervention. As the intervention was successfully implemented by teachers in a large number of classrooms, this study suggests that Tekster is a promising approach for improving students' writing in general education.

Keywords: writing, writing instruction, observational learning, strategy instruction, elementary grades

Supplemental materials: http://dx.doi.org/10.1037/edu0000206.supp

Despite the fact that writing plays an important role in academic and career success, research shows large numbers of students from many different countries fail to develop essential writing skills (e.g., Department for Education, 2012; Salahu-Din, Persky, & Miller, 2008). For example, a recent national assessment in the Netherlands revealed most elementary-aged students were unable to write texts that convey a single, simple message to the reader and students' writing skills improved negligibly from fourth to sixth grade (Kuhlemeier, Til, Hemker, de Klijn, & Feenstra, 2013). Furthermore, the Dutch Inspectorate of Education determined the quality of writing instruction to be sufficient in only one third of the nation's schools (Henkens, 2010). Thus, an improvement in elementary-level writing instruction in the Netherlands is required. For this purpose, we developed the writing program Tekster. Tekster incorporates several research-supported instructional practices and addresses both the focus and mode of instruction (what we teach and how we teach it). The effectiveness of Tekster for fourth-, fifth-, and sixth-grade Dutch students was tested in this study.

Focus of Instruction

The major problem developing writers face is cognitive overload. Writers have to perform several resource-demanding cognitive activities simultaneously, such as activating prior knowledge, generating content, planning, formulating, and revising—all while taking into account the communicative goal of the text and the intended audience (Fayol, 1999). The amount of attention required for foundational skills (e.g., handwriting, spelling, sentence, and paragraph construction) also needs to be considered with developing writers because they often lack automaticity in these areas (McCutchen, 2011). Developing writers predominantly use a "knowledge-telling" approach to overcome cognitive overload. That is, they write whatever happens to come to mind and typically

This article was published Online First April 13, 2017.

Renske Bouwer, Monica Koster, and Huub van den Bergh, Utrecht Institute of Linguistics, Utrecht University.

This research was supported by grant 411-11-859 from the Netherlands Organization for Scientific Research (NWO). We would like to thank our editor, Tanya Santagelo, for her support, encouragement and constructive feedback during the reviewing process.

Correspondence concerning this article should be addressed to Renske Bouwer, who is now at the Department of Training and Educational Sciences, University of Antwerp, Gratiekapelstraat 10, 2000 Antwerpen, Belgium. E-mail: Renske.Bouwer@uantwerpen.be

focus only on the content of their texts (Bereiter, Burtis, & Scardamalia, 1988). With this approach text production is restrained by the storage and retrieval capacity of short-term memory (STM; Miller, 1956) and this often results in texts that are not sufficiently adapted to the communicative goal and intended audience (Berninger et al., 1992; McCutchen, 1996). To improve students' writing performance, instruction should be aimed at helping them develop the knowledge and skills required to manage the cognitive overload that often occurs when composing.

Strategy Instruction

An effective way to help developing writers manage cognitive overload is to teach them to use strategies that reduce the number of cognitive processes that are active at the same time (Kellogg, 1988, 2008). For instance, when students are taught to plan during the prewriting phase, they can focus on other processes while drafting. A substantial body of research has examined the impact of explicitly teaching students to use writing strategies. Some studies investigated strategies designed to guide general writing processes, such as brainstorming (Troia & Graham, 2002) or revising (Fitzgerald & Markham, 1987), whereas others featured genre or task specific strategies, such as writing a narrative text (Brunstein & Glaser, 2011) or a persuasive essay (Wong, Hoskyn, Jai, Ellis, & Watson, 2008). Despite the diversity of research examining explicit strategy instruction, results are remarkably consistent and positive. For example, several recent meta-analyses reported large average weighted effect sizes (ESs), ranging from 0.82 to 1.15, for explicit strategy instruction (Graham, 2006; Graham, McKeown, Kiuhara, & Harris, 2012; Graham & Perin, 2007; Hillocks, 1984; Koster, Tribushinina, De Jong, & Van den Bergh, 2015).

Self-Regulation

When explicit strategy instruction is combined with teaching self-regulatory skills, the impact on students' writing is even greater (Graham et al., 2012). *Self-regulation* is "the process whereby individuals activate and sustain behaviors, cognitions, and affect, which are systematically oriented toward the attainment of goals" (Schunk, 2012, p. 123). Essential self-regulatory skills in writing include setting communicative, process, and progress goals, and subsequently monitoring progress toward those goals (Flower & Hayes, 1981). The most prominent and well-researched model for explicitly teaching writing strategies and self-regulation is self-regulated strategy development (SRSD; Harris, Graham, Mason, & Saddler, 2002). SRSD has been validated through research spanning over three decades and involving a wide range of students in many different instructional environments. Results from SRSD studies consistently show the approach is highly effective for improving students' writing performance (ES = 1.17, Graham et al., 2012).

Students' self-regulation is positively affected by the attainment of specific goals which, in turn, enhances self-efficacy for writing (Latham & Locke, 1991; Schunk, 1990). Students benefit the most from challenging, but attainable, goals that specify what needs to be accomplished through the writing task (Schunk, 1990). For example, previous research shows assigning students specific goals for improving the content of their texts and making them aware of the intended audience leads to improvements in planning, drafting, and revising (Ferretti, Lewis, & Andrews-Weckerly, 2009; Ferretti, MacArthur, & Dowdy, 2000; Graham, MacArthur, & Schwartz, 1995; Midgette, Haria, & MacArthur, 2008). Research also indicates short-term writing goals are more beneficial than goals spanning longer periods of time (Latham & Locke, 1991).

Text Structure Instruction

To be proficient writers, students need to be able to establish their own composing goals for different writing tasks. They also need to know how to create texts that meet the goals they set (Schoonen & De Glopper, 1996). Explicit text structure instruction, whereby the elements and organization of different text types are specifically taught, has been shown to help students acquire the knowledge needed to set and achieve writing goals. Research examining the impact of explicit text structure instruction for elementary-aged students spans three major genres: narrative (Fitzgerald & Teasley, 1986; Gordon & Braun, 1986), persuasive (Crowhurst, 1990, 1991; Scardamalia & Paris, 1985), and informative (Bean & Steenwyk, 1984; Raphael & Kirschner, 1985). The findings from two recent meta-analyses provide further support for the positive effect of text structure instruction. Graham et al. (2012) and Koster et al. (2015) reported an average weighted ES for explicit text structure instruction of 0.59 and 0.76, respectively.

Mode of Instruction

For developing writers, learning to write and task execution are often inextricably linked. Simultaneously, students have to learn how to write and produce texts (Rijlaarsdam & Couzijn, 2000). However, because text production is so cognitively demanding for developing writers, this instructional approach often results in students having minimal attentional capacity left to learn from their writing experiences (Rijlaarsdam & Couzijn, 2000). Thus, to optimize the way writing is taught, it is important to carefully consider the format and sequence of instruction.

Observational Learning

One way to separate task performance from learning is to provide opportunities for observation (Zimmerman & Risemberg, 1997). Observing someone complete an unfamiliar task is less demanding on working memory than having to actually perform the task yourself. This is particularly true when the skill being learned is cognitively complex—such as writing (Rijlaarsdam, 2005). Observational learning was first described and studied by Bandura (1986) as part of social–cognitive learning theory. Within this framework, observation allows individuals to gain insight into the usefulness and consequences of the behavior being modeled. Behavior that is evaluated positively and considered useful will be retained (Schunk, 2012). Observational learning can be applied to teaching writing in two ways: through different types of modeling (before and during writing) and through reader feedback (during and after writing).

Teacher modeling. In writing instruction, observational learning is frequently implemented by means of teacher modeling.

Modeling involves explaining, demonstrating, and verbalizing one's thoughts and actions, with the aim of eliciting behavioral change in an observer (Schunk, 2012). This kind of modeling prepares students for the forthcoming composing task in the initial phase of the writing process. Several studies have demonstrated the effectiveness of teacher modeling as an instructional practice for teaching writing strategies (e.g., Fidalgo, Torrance, Rijlaarsdam, Van den Bergh, & Lourdes Álvarez, 2015; Graham, Harris, & Mason, 2005).

Mastery versus coping models. Models can show either mastery or coping behavior. Mastery models demonstrate a flawless performance, whereas coping models display common challenges, as well as ways to overcome those difficulties and gradually improve performance (Schunk, 1987; Zimmerman & Kitsantas, 2002). In a study on revision skills, Zimmerman and Kitsantas (2002) found observing a coping model raised students' self-efficacy and enhanced their performance more effectively than a mastery model. Research suggests observing coping models is particularly beneficial for weaker students. This may be the result of explicitly seeing how to overcome difficulties and/or watching someone who is perceived as similar improve performance over time (Schunk, 1987).

Peer modeling. When peers—rather than teachers—act as models, perceived model-observer similarity is even higher because of the developmental resemblance (Schunk, 1987). Peer modeling has been investigated in several studies. Raedts, Rijlaarsdam, Van Waes, and Daems (2007) found observing video-based peer models improved text organization and self-perception of writing performance. Couzijn (1999) demonstrated observing peer models can have large effects on argumentative text-writing. Van Steendam, Rijlaarsdam, Van den Bergh, and Sercu (2014) found both more and less proficient writers benefited from peer modeling during a collaborative revising task. Braaksma (2002) and Braaksma, Rijlaarsdam, Van den Bergh, and Van Hout-Wolters (2004) found observing peer models positively impacted students' writing performance and writing processes. Braaksma's (2002) findings also provided support for the model-observer similarity hypothesis. Weaker students performed better after focusing on a weaker peer model, whereas stronger students showed greater improvement after focusing on a stronger peer model. Observing mastery peer models may be especially beneficial for stronger students because they set positive standards for performance (Zimmerman & Kitsantas, 2002). In contrast, observing coping peer models may be especially effective for weaker students, as they enhance self-efficacy and motivation (Schunk, 1987). It should be noted, however, all the aforementioned peer modeling studies were conducted with (post)secondary students, rather than students in the elementary grades.

Reader reaction. Whereas teacher and peer modeling primarily focus on teaching students aspects of the writing process, a different form of observational learning can be used to provide students with feedback on the communicative effectiveness of their compositions. In contrast with oral communication, separation in time and space results in writers rarely receiving any direct cues or feedback from those who read their text (Rijlaarsdam et al., 2008). This can be particularly disadvantageous for developing writers who are not yet proficient in self-evaluation. Observational learning can help bridge this gap and develop students' understanding of how readers experience and perceive their texts (Couzijn & Rijlaarsdam, 2004; Schriver, 1992). Several researchers (Couzijn, 1995; Couzijn & Rijlaarsdam, 2004; Holliway & McCutchen, 2004; Rijlaarsdam, Couzijn, Janssen, Braaksma, & Kieft, 2006) have shown students' writing can improve after observing the effect their text has on readers. Meta-analytic results indicate both feedback and peer interaction can enhance writing quality. The average weighted ESs reported by Graham et al. (2012) and Koster et al. (2015) for feedback were 0.80 and 0.88, and for peer interaction were 0.89 and 0.59, respectively.

Gradual Release of Responsibility

Improving students' writing performance cannot be accomplished solely through observational learning; there comes a time when students need to transition from observing writing models to actually composing themselves. Moreover, to successfully complete a writing task, students must eventually progress through all the stages of the writing process. One way to ease the transition between observation and task execution is through the gradual release of responsibility (Pearson & Gallagher, 1983). With this approach, cognitive load is gradually shifted from observing models, to guided practice, and finally to independent performance. The gradual release of responsibility model builds on Vygotsky's (1980) sociocultural theory and concept of the zone of proximal development. Vygotsky defined the zone of proximal development as the area between a student's level of independent performance and potential development, as determined by assisted performance. Teachers can facilitate progression from assisted to independent performance through scaffolding. That is, they control elements of a task initially beyond a student's capacity to enable the development of skills within the range of competence (Wood, Bruner, & Ross, 1976). As a student progresses, teacher assistance is gradually reduced. For scaffolding to be successful, teachers need to help students develop strategies that are transferrable to new tasks and situations (Bodrova & Leong, 1998).

Writing instruction programs that use gradual release of responsibility and scaffolding techniques have been shown to improve students' written language skills (Graham et al., 2005; Graham et al., 1995). Many of these programs also use explicit instruction to activate students' background knowledge and help them understand the purpose and benefits of the strategy being taught. For upper-elementary aged students, generalization of strategy use to other tasks and domains is promoted through comprehensive and explicit instruction regarding how and when a strategy can best be applied (O'Sullivan & Pressley, 1984).

Aim of the Study

The main purpose of this study was to test the effectiveness of Tekster (Koster, Bouwer, & Van den Bergh, 2014a, 2014b, 2014c), a comprehensive writing instruction program we developed to be implemented by fourth, fifth, and sixth grade general education teachers in the Netherlands. The main focus of Tekster is teaching students a general writing strategy, as well as the self-regulation skills needed to use the strategy successfully. Genre-specific features are addressed through explicit instruction in text structure. The predominant mode of instruction is observational learning, complemented by explicit instruction and guided practice that includes extensive scaffolding and the gradual release of responsibility (Wood et al., 1976). In this regard, Tekster bears close resemblance to SRSD (Harris et al., 2002) and cognitive self-regulation instruction (Fidalgo et al., 2015).

In the present study, we investigated whether Tekster improved the quality of writing produced by fourth, fifth, and sixth grade Dutch students and whether the effect of the intervention was maintained over time. In addition, we examined whether the effect differed based on students' grade level, gender, or level of writing proficiency.

Method

Sample

Seventy-six upper-elementary teachers, representing 60 classrooms, volunteered to participate in the study. The majority of teachers were female (82%) and all participants held the required professional certification. The study took place in 27 schools, located throughout the Netherlands. Eleven schools were in the northern part of the country, nine were centrally located, and seven were in the southern region. Sixty percent of the schools were religiously affiliated (11 Catholic, two Protestant, two Reformed, one Islamic) and 40% were public. Ten schools had one participating classroom, whereas two to five classrooms participated in the other 17 schools. With regard to grade level, there were 20 fourth-grade classes, 13 fifth-grade classes, 16 sixth-grade classes, and 11 multigrade classes (i.e., a combination of two or three grade levels). The average number of students per class was 23.6 (*SD* = 5.6), half of whom were female. The schools, teachers, and students in our sample did not differ significantly from the Dutch population in terms of denomination (Ministry of Education, Culture, & Science, 2015), gender (Central Office for Statistics, 2015; Inspectorate of Education, 2012), or classroom size (Central Office for Statistics, 2015).

In total, 1,420 students participated in the study: 477 fourth graders (*M* age = 9.40, *SD* = 0.62), 454 fifth graders (*M* age = 10.40, *SD* = 0.61), and 489 sixth graders (*M* age = 11.50, *SD* = 0.64).[1] A small number of individual students dropped out because they changed schools during the study. Specifically, 17 students (1.2%) completed only the pretest measures and 37 students (2.6%) completed only one of the two posttest measures.

Design of the Study

To analyze whether Tekster improved students' writing quality, we used a switching replication design (Shadish, Cook, & Campbell, 2002) with two groups and three measurement occasions (M1, M2, M3; see Table 1). In the first phase of the study, from M1 to M2, teachers and students in Group 1 worked with Tekster—instead of their regular writing instruction program—for 8 weeks, completing two lessons per week. Group 2 served as a control group during this period; teachers and students continued with their existing writing activities and routines. During the second 8-week phase, between M2 and M3, the intervention switched between groups. Group 2 implemented Tekster and Group 1 returned to their original writing program. M3 served as a posttest for students in Group 2, as well as a delayed posttest for students in Group 1, which enabled us to measure their level of retention.

A switching replication design is superior to a regular pretest posttest (quasi-) experimental design because the intervention is implemented in both groups, but during different time intervals (Shadish et al., 2002). It is not only a more ethical design, as all students eventually benefit from the intervention, but it also allows for a test of internal validity. If the intervention is equally effective in both groups, the effect does not likely depend on characteristics of a particular group. If the effect of the intervention is not equally effective in both groups, internal validity might be threatened. Moreover, because the intervention is replicated in two groups, important information about the reproducibility and generalizability of the results is generated (Open Science Collaboration, 2015). The design also provides information about maintenance effects because it includes a delayed posttest (M3) for students in Group 1.

Assignment of schools to groups. The school holiday calendar determined which schools were assigned to Group 1 and Group 2. Specifically, schools located in the northern region were assigned to Group 1 and those in the south were assigned to Group 2. Schools from the middle region were randomly assigned to Group 1 or 2. Group 1 included 14 schools, 31 teachers (84% female), and 29 classes. Group 2 included 13 schools, 45 teachers (80% female), and 31 classes. Table 2 contains a summary of student information for each group. The number of students per grade was similar for both groups, $\chi^2(2) = 2.67$, $p = .26$, and there were no statistically significant differences in gender distribution, $\chi^2(1) = 2.21$, $p = .14$, or age, $t(1414) = -1.31$, $p = .19$, between groups.

Writing Instruction

Existing instruction. The intervention program was compared to the existing writing instruction practices used in each participating classroom. In the Netherlands, writing is traditionally taught as part of the Dutch language curriculum. According to a report published by the Dutch Inspectorate of Education (Henkens, 2010), of the 8 hours per week reserved for language teaching, an average of only 45 min is devoted to writing. Writing lessons are primarily product-focused: Students receive minimal support during the writing process and are not taught how to approach writing tasks. In addition, in the majority of schools, students' writing performance is not monitored and they are rarely given feedback on their compositions. Many of the Inspectorate's findings were recently corroborated by a study exploring how writing is taught by 51 Dutch elementary teachers (Rietdijk, Van Weijen, Janssen, Van den Bergh, & Rijlaarsdam, 2015). For example, 94% of the teachers said they spend less than one hour per week teaching writing. They also described typical writing lessons as dominated by independent student work, with only one third of the time being used for plenary instruction. Modeling, individualized support, and providing students with feedback were all reported to be uncommon practices. In contrast with what was reported by the Inspectorate, however, teachers who participated in Rietdijk et al.'s study

[1] Specific information on students' special educational needs was not available. Typically, in an average Dutch general education classroom, 20% to 25% of the students will have learning and/or behavioral difficulties that require additional, individualized attention (Koopman, Ledoux, Karssen, Van der Meijden, & Petit, 2015).

Table 1
Design of the Study

Group	Pretest (M1)	Phase 1 (8 weeks)	Posttest (M2)	Phase 2 (8 weeks)	Delayed posttest (M3)
1	Tasks a, b, c	Tekster intervention	Tasks d, e, f	Existing writing instruction	Tasks g, h, i
2		Existing writing instruction		Tekster intervention	

said they do attend to the different stages of the writing process. For example, a majority of respondents reported using prewriting activities and half said they ask students to revise their texts.

Tekster. The intervention program, Tekster, included a series of 16 grade-level specific lessons, compiled in a student workbook and accompanied by a teacher's manual (Koster et al., 2014a, 2014b, 2014c). Tekster incorporates several different research-based practices to address both the focus and the mode of instruction. Table 3 gives an overview of how the program's three design principles—writing strategies, text structure, and self-regulation skills (see Rijlaarsdam, Janssen, Rietdijk, & Van Weijen, in press)—were operationalized into specific teaching and learning activities.

Lesson format and writing strategies. Tekster lessons followed a generally consistent format, with each lesson typically lasting between 45 and 60 min (see Table 4). The focal point of instruction was the writing strategy students learned to help guide them through the steps of the writing process. A mnemonic device was used to help students remember and apply the writing strategy: The first letter of each strategy step formed an acronym that spelled the name of an animal.

- Grade 4 students learned VOS *(which means fox*): Verzinnen (*generate content*), Ordenen (*organize*), Schrijven (*write*).
- Grade 5 students learned DODO (which means *dodo*): Denken (*think*), Ordenen (*organize*), Doen (*do*), Overlezen (*read*).
- Grade 6 students learned EKSTER (which means *magpie*): Eerst nadenken (*think first*), Kiezen & ordenen (*choose & organize*), Schrijven (*write*), Teruglezen (*reread*), Evalueren (*evaluate*), Reviseren (*revise*).

The three animals were used as a common theme for all the lessons in the corresponding grade level and small images representing the animals provided additional visual support. A sample Tekster lesson is included as Supplemental Appendix A in the online supplemental materials.

Lesson content and sequence. During the first Tekster lesson, students were introduced to the acronym animal corresponding with the writing strategy they would learn, through a story. They also practiced the steps of the strategy for the first time. In subsequent lessons, students learned to apply the writing strategy to different types of texts. All the practice writing tasks were authentic and represented a variety of communicative goals and audiences. For instance, students in each grade wrote texts that were descriptive (e.g., personal advertisement, self-portrait), narrative (e.g., story for kindergartener, newspaper article), persuasive (e.g., email nominating for a TV program, flyer recruiting new members for a club), instructive (e.g., recipe, rules for a game) and personal communications (e.g., holiday postcard, party invitation). The writing tasks for each grade level were in line with the Dutch Ministry of Education's goal for students at the end of elementary school "to write coherent texts, with a simple linear structure on various familiar topics; the text includes an introduction, body, and ending" (Expert Group Learning Trajectories, 2009, p.15).

Table 2
Student Characteristics

	Group 1			Group 2		
Grade	*n*	% female	Mean age (*SD*)	*n*	% female	Mean age (*SD*)
4	245	47%	9.41 (.58)	232	54%	9.39 (.65)
5	217	51%	10.39 (.63)	237	54%	10.42 (.59)
6	226	46%	11.50 (.67)	263	48%	11.50 (.62)
Total	688	48%	10.41 (1.07)	732	52%	10.48 (1.07)

The level of difficulty for the writing tasks ascended through the grades as follows: In Grade 4, tasks featured an intended audience in close proximity to the student, such as classmates, friends, and (grand)parents. In Grade 5, the target audience expanded to include people with whom students had a more distal relationship but yet were still familiar, such as teachers, relatives, and neighbors. In Grade 6, students also wrote texts intended for unfamiliar people, such as a newspaper editor and owner of a company.

Lesson development. Tekster lessons were developed in close collaboration with 16 elementary school teachers. These teachers were divided into three design teams that met once a month over a period of six months. After receiving an introduction to the program's guiding principles, two design teams worked on developing the practice writing tasks that would eventually be integrated into Tekster lessons. Each writing task needed to focus on a topic of interest to upper-elementary students and have a clearly specified communicative goal and target audience. Teachers piloted the writing tasks with their own students and received feedback from their team members and the authors during the monthly meetings. The third design team made peer modeling video clips that were used as part of Tekster instruction. After the writing tasks and video clips were created, the authors wrote the detailed lesson plans for each grade level and subsequently piloted the program (see Koster, Bouwer, & Van den Bergh, 2016).

Teacher training. The teachers who participated in this study learned about Tekster during a 4-hr session training session led by the authors and held in small groups consisting of no more than 12 people. At the beginning of the training session, each teacher received a Tekster teacher's manual that was divided into two sections. The first section included an introduction to the program (e.g., goals, guiding principles) and descriptions of the essential components (e.g., instructional model, general lesson format and sequence, specific research-based practices). An overview of the

Table 3
Overview of Design Principles, Learning and Teaching Activities of Tekster

Design principles		Tekster intervention	
Focus of instruction	Mode of instruction	Learning activities	Teaching activities
1. Writing strategies	a. Observational learning	Observe/discuss/compare model(s) (teacher or peer) and apply the writing strategy in different stages of the writing process	Model strategy use by thinking aloud while performing (part of) the writing task
	b. Explicit instruction	Listen actively, retrieve relevant background knowledge from memory, take notes	Explain the components of the strategy, make students aware of the purpose and benefits of using a writing strategy, activate students' background knowledge
	c. (Guided) practice	Apply the steps of the strategy to authentic writing tasks in various genres with clear communicative goals and intended audience	Provide help when needed through scaffolding and process feedback
2. Text structures	a. Observational learning	Before writing: Observe/discuss/compare model(s), (teacher or peer) talking about criteria for various text types, compare and discuss model texts of the same text type to derive criteria and conventions for a good text	Before writing: Model the relevant aspects of the text type, provide model texts or show video clips of peer modeling
		After writing: Evaluate peer/own text on the basis of the previously discussed criteria and give feedback (reader reaction), observe reader reaction, observe model revising on the basis of feedback	After writing: Evaluate students' texts on the basis of previously discussed criteria, give feedback (reader reaction), model how to revise the text
	b. Explicit instruction	Listen actively, take notes	Explain why and how the criteria and conventions should be used, discuss important criteria and conventions using model texts
	c. (Guided) practice	Apply discussed criteria to authentic tasks in various genres with clear communicative goals and intended audience	Provide help when needed through scaffolding and product feedback
		After writing: Give peer feedback and assess own text according previously discussed criteria	
3. Self-regulation skills	a. Observational learning	Observe/discuss/compare model(s), (teacher or peer) setting goals and monitoring progress in relation to goals during the writing process, observe/discuss/compare effect of self-regulation on the written product	Model self-regulation during writing, setting a goal for writing and monitoring progress towards this goal
	b. Explicit instruction	Listen actively, take notes	Explain the differences between various communicative goals, explain the importance of setting communicative goals for writing in advance, and show when and how during the writing process progress towards the communicative goal can best be monitored
	c. (Guided) practice	Set communicative goal before writing, monitor progress towards this goal during writing, regulate own writing process and adapt if necessary, evaluate written product in relation to communicative goal, revise if necessary.	Provide help when needed through scaffolding, and self-regulation feedback

Table 4
Tekster's General Lesson Format

Lesson phase	Learning and teaching activities
1	Goal of the lesson is explicitly stated (**3b**)
2	Plenary introduction in which specific characteristics of text type are addressed through modeling (**2a**), comparing model texts (**2a**), or explicit teacher instruction (**2b**)
3	Introduction of authentic writing assignment in which communicative goal and intended audience are explicated (**3b**)
4	Acronym for the strategy is explicitly named (**1b**)
5	Content is generated in keywords (first step of the strategy; gradual release of responsibility from **1a** to **1c**, **3a** to **3c**)
6	Content is generated in keywords (second step of the strategy; gradual release of responsibility from **1a** to **1c**, **3a** to **3c**)
7	Text is written using organized content (third step of the strategy; **1c**, **2c**, **3c**)
8[a]	Students' texts are read (fourth step of the strategy; **2a**)
9[a]	Students' texts are evaluated by answering evaluative questions and/or giving feedback (fifth step of the strategy; **2a**)
10[b]	Students' texts are revised on the basis of the received feedback (sixth step of the strategy; **3c**)

Note. Bold numbers refer to focus and mode of instruction as shown in Table 3.
[a] Only for Grades 5 and 6. [b] Only for Grade 6.

study was also provided. The second section of the teacher's manual contained the 16 lesson plans teachers were cxpected to implement during the 8-week intervention period. A DVD with peer modeling video clips and examples of teacher modeling was also provided.

The Tekster teacher's manual served as a guide during the training session. First, teachers learned about the program's theoretical framework, goals, and general structure. Then, they focused on specific instructional practices and effective lesson implementation. For instance, one characteristic of effective teacher feedback about writing is providing students with individualized comments, based on their areas of strength and need (Parr & Timperley, 2010). Therefore, during the training session, teachers learned about and collaboratively practiced the underlying skills needed to provide this type of feedback (i.e., accurately assessing the quality of students' texts and adapting comments accordingly). At the end of the training session, the authors stressed that it was very important for each teacher to carefully read the entire teacher's manual and watch the full DVD before implementing Tekster.

Intervention Fidelity

Several fidelity measures were used to determine whether teachers implemented Tekster as intended. Specifically, fidelity was operationalized three ways: number of lessons taught by each teacher, number of lessons completed by each student, and teachers' adherence to the lesson plans included in the teacher's manual. The strategies used to collect fidelity data included reviewing teachers' log books, reviewing students' workbooks, and observing classroom instruction.

Teacher logbooks. Each teacher was asked to maintain a logbook during Tekster implementation to document the number of lessons completed and the duration of each. After the intervention period, 75% of the logbooks were returned. Analysis of those data indicated teachers taught an average of 10 (out of the intended 16) Tekster lessons. The average number of minutes required to complete a lesson was 43, with a range of 29 to 58.

Student workbooks. We collected and reviewed students' workbooks after the intervention period to determine the number of lessons each student completed. A lesson was considered complete if a student's workbook contained a text corresponding with the practice writing task for that lesson. Analysis of these data revealed considerable variability in the number of lessons students completed. On average, students completed 10 lessons (*SD* = 4); however, 8% of students completed less than four lessons and 53% of students completed at least 10 lessons.

Classroom observations. Observations were conducted in two thirds of the classrooms (selected at random) in Group 1 and Group 2. Each observation lasted the full length of the lesson and took place approximately half-way through the intervention period. The observations for each group occurred over a 2-week period; thus, there was variation in the particular lessons observed. Ten trained undergraduate students served as observers in this study. Because each classroom was observed by only one person, the reliability of the observational data was not able to be assessed.

Our observation instrument was based on the work of Hintze, Volpe, and Shapiro (2002) and designed to gather two types of data: general adherence to the lesson plan and frequency of using two key instructional practices—teacher modeling and the writing strategy. To assess whether a lesson plan was being implemented as intended, every 20 s observers tallied whether a teacher was on task (i.e., executing the actions specified in the lesson plan for that phase of instruction) or off task (i.e., doing something unrelated to writing instruction). Each on task tally was further categorized as plenary (i.e., involving the whole class) or individualized (i.e., involving individual students or a small group of students). To measure the frequency of teacher modeling and strategy use, observers recorded the number of times a teacher modeled something for the class and the number of the times a teacher referred to the writing strategy acronym or steps.

Analysis of the observational data indicated teachers adhered closely to what was specified in the Tekster lesson plans. On average, teachers were on task 92% of the observed instructional time and their actions were consistent with the general framework and key elements of Tekster. For example, the division between plenary and individualized instruction was relatively equal (on average, 54% and 46%, respectively), as intended. Teacher modeling and use of the writing strategy were also evident (on average, 1.3 and 1.4 times per lesson, respectively).

Assessment of Writing Quality

Writing tasks. Because generalization of writing proficiency across genres is not warranted when scores are obtained with only one writing task (Bouwer, Béguin, Sanders, & Van den Bergh, 2015), we assessed students' writing skills at each measurement occasion using three different types of texts: descriptive (tasks a, d, g), narrative (tasks b, e, h) and persuasive (tasks c, f, i), as shown in Table 1. The three tasks for each genre were as similar as possible and differed only in topic, not format. All nine writing tasks were developed by the authors for the purpose of this study and in conjunction with other experts in the field. To increase the likelihood students would produce texts of reasonable length, specific attention was given to ensure an appropriate level of difficulty and topical interest. For each task, students received a handout that included the written prompt, topically related image, and space for prewriting (if desired). A sample prompt for each text type is provided as Supplemental Appendix B in the online supplemental materials.

Administration of writing tasks. The writing tasks used as assessments in this study were administered by the participating teachers to students in their classroom during regular instructional time. Teachers were asked to administer the three writing tasks for each measurement occasion within one week, but not on the same day. Students completed each writing task independently and without a time constraint. Teachers were instructed not to provide students with any additional assistance while they completed the assessments.

Rating Writing Quality

We anonymized all student texts to reduce the likelihood characteristics such as gender or grade level would influence raters' judgments (Peterson, 2000). However, due to the scope of this study (1,420 students and nine writing tasks, resulting in approximately 12,780 written texts), it was not feasible to type students' handwritten work as a way to control for possible presentation effects (Graham, Harris, & Hebert, 2011). Global text quality was assessed using a continuous (interval) rating scale with five benchmarks (Blok & Hoeksma, 1984; Bouwer, Koster, & Van den Bergh, 2016). The midpoint on the scale was an average quality text, assigned an arbitrary score of 100. The other four benchmark texts were located one and two standard deviations above and below the midpoint and scored (in ascending order) as 70, 85, 115, and 130. A different benchmark scale was constructed for each text type. Supplemental Appendix C in the online supplemental materials contains a sample benchmark rating scale.

The rating scale benchmarks originated from a preliminary investigation of a randomly selected subsample drawn from all the texts (i.e., all three grade levels and genres) written during M1. Five experienced Grade 4–6 teachers rated the subsample holistically and their scores were averaged. Benchmarks were then selected based on two criteria: (a) the text was a good representation of the quality level (-2 *SD*, -1 *SD*, 0, $+1$ *SD*, $+2$ *SD*); and (b) the level of rater agreement about text quality was high.

The raters for the full assessment sample were also experienced Grade 4–6 teachers ($n = 47$). Raters were trained in advance how to use the benchmark scales and they were blind to experimental conditions. Each rater compared a student's text to the benchmarks and assigned a score, accordingly. Each text was rated by a jury of three people, using a design of overlapping rater teams. With this method, all the student texts were divided randomly into subsamples, equal to the number of raters. Each rater received three subsamples, based on a predetermined design. The overlap in subsamples allowed us to approximate the reliability of raters and juries (Van den Bergh & Eiting, 1989). The average reliability of jury ratings across tasks was high, $\rho = .89$, with the variation between tasks being $\rho = .86$–.91. The final quality score for each text was determined by computing the mean of the three ratings. The raters' scores were normalized for each task using Blom's rank-based normalization formula (see Solomon & Sawilowsky, 2009) because they appeared to be negatively skewed (i.e., low quality texts tended to be scored more extremely).

Data Analyses

The data in this study were hierarchically organized; scores were cross-classified with students and tasks, and students were nested within classes. Therefore, the data were analyzed by applying different (cross-classified) multilevel models in which parameters were added systematically to the model. In such models, all students—including those with partly missing values—are taken into account.

The effectiveness of Tekster across groups and grade levels was tested with six models. Model 1 was the basic null model in which we only accounted for random error (S^2_e) and random effects of students (S^2_s), tasks (S^2_t), and classes (S^2_c). That is, writing scores were allowed to vary within and between students, between tasks (including systematic variation due to genre), and between classes. In Model 2, measurement occasion was added as a fixed effect to test whether average scores differed over time. Whether the variances within and between students, and between classes, differed between the three measurement occasions was tested in Model 3. In Model 4, group was added as a fixed effect to test whether average scores differed between the two groups. Model 5 tested the main effect of the intervention by estimating the interaction between group and measurement occasion. This model included the restriction that the effect of the intervention was the same in the two groups. Finally, this restriction was removed in Model 6 to test whether the intervention was equally effective in Group 1 and 2 which, in essence, provided a check on the internal validity of the experiment.

The maintenance effect of the intervention was tested by performing a specific contrast analysis of students in Group 1. In this analysis, students' posttest and delayed posttest scores were compared. To test whether the intervention was equally effective across grade levels, we applied two additional models. In the first model, grade was added as a fixed effect to test whether average scores differed between the three grades. In the second model, the interaction effect between the intervention (Measurement Occasion × Group) and grade level was added to test whether the intervention was equally effective across the three grades.

The role of gender on the effectiveness of the intervention was tested by two additional models. In the first model, gender was added as a fixed effect to test whether average scores differed between male and female students. In the second model, the interaction effect between the intervention (Measurement Occasion × Group) and gender was added to test whether the intervention was equally effective for males and females.

To test whether the intervention was equally effective for students with different levels of writing proficiency, we performed an aptitude treatment interaction analysis. For this analysis, the regression of students' pretest scores on their posttest outcomes was estimated per group.

Results

Effect of the Intervention

Results of the fit and comparison of the six models are shown in Table 5. There was a fixed effect of measurement occasion—Model 2 vs. Model 1, $\chi^2(2) = 279.61$, $p < .001$—which indicates average writing scores were not equal over time. Allowing the variances to differ between measurement occasions significantly improved the model—Model 3 vs. Model 2, $\chi^2(12) = 657.61$, $p < .001$; thus, for at least one level (students, tasks, classes, and/or random error), the variance was not homogeneous across measurement occasions. The main effect for group—Model 4 vs. Model 3, $\chi^2(1) = 1.32$, $p = .25$—was not statistically significant, indicating average scores were the same for students in Group 1 and 2.

There was a statistically significant effect for the intervention—Model 5 vs. Model 4, $\chi^2(1) = 24.98$, $p < .001$—as indicated by the interaction between group and measurement occasion. That is, differences in scores measured at two occasions (i.e., first and second or second and third) were not the same for students in the intervention and control conditions. The effect of the intervention on differences in scores appeared to be the same for students in Group 1 and 2—Model 6 vs. Model 5, $\chi^2(1) = 0.12$, p = .73. To verify the interaction effect between group and measurement occasion, we tested two additional contrasts. The interaction between group and the first two measurement occasions was statistically significant, $\chi^2(1) = 11.52$; $p < .001$; the difference in mean scores between measurement occasions was larger for Group 1. The interaction effect between group and the latter two measurement occasions was also statistically significant, $\chi^2(1) = 30.86$; $p < .001$; the difference in mean scores between the second and third measurement occasion was larger for Group 2.

Parameter estimates of Model 5 are summarized in Table 6 and a graphical display of the intervention effect is presented in Figure 1. The variance within and between students decreased over time, as did the variance between classrooms. The decrease in between-class variance means classes became more homogeneous over time. The reduction in within-student variance resulted from smaller interaction effects between students and tasks, indicating students' writing also became more homogeneous.

To estimate the magnitude of Tekster's effect, we compared the impact of the intervention to the total variance (Cohen's *d*). The overall ES (i.e., across all students, teachers, and tasks; based on the mean number of student-completed lessons) was 0.32. Because we discovered considerable variability in the number of lessons students completed ($M = 10$ and $SD = 4$, as reported above under the Intervention Fidelity section), we included this variable as a fixed factor in the analyses. The results indicated a statistically significant, positive relationship between the number of student-completed lessons and the intervention effect, $\beta = 0.21$ ($SE = 0.09$, $p < .01$). On average, students who completed all 16 Tekster lessons had a gain score of 5.99, which translates to an ES of 0.40.

Maintenance. For students in Group 1, the impact of Tekster was measured immediately after the intervention period (M2) and again, 8 weeks later (M3). Results of the specific contrast analyses indicated the effect of the intervention was maintained over time. There was a statistically significant increase in students' scores between M1 and M3, $\chi^2(1) = 23.14$, $p < .001$, but the difference between M2 and M3 was not statistically significant, $\chi^2(1) = 2.06$, $p = .15$.

Grade level. The main effect for grade level was statistically significant, $\chi^2(2) = 54.40$, $p < .001$, meaning average scores differed for students in Grade 4, 5, and 6. The interaction between the intervention and grade level was also statistically significant, $\chi^2(2) = 14.21$, $p < .001$, indicating the impact of Tekster differed based on grade level. On average, Grade 4 students' scores increased by 4.86 points (ES = 0.34), Grade 5 students' scores increased by 5.00 points (ES = 0.35), and Grade 6 students' scores increased by 4.23 points (ES = 0.30). A graphical display of the intervention effect for each grade level is presented in Figure 2.

Gender. The main effect for gender was statistically significant, $\chi^2(1) = 319.70$, $p < .001$. On average, female students' scores exceeded male students' scores by 7.62 points. The effect of Tekster, however, was not gender dependent, as indicated by a nonsignificant improvement in the model when the interaction between group, measurement occasion, and gender was allowed, $\chi^2(1) = 0.10$, $p = .75$.

Writing proficiency. For students in Group 1, who participated in the intervention in the first 8 weeks, the regression coefficient of the scores of the first measurement occasion on the second measurement occasion equaled 0.60 ($SE = 0.02$). The regression coefficient for Group 2 in the same period equaled 0.59 ($SE = 0.03$), which was a nonsignificant difference ($t = 0.20$; $p = .42$). Hence, the results did not show an aptitude treatment interaction, indicating the effects of the intervention did not depend on students' writing proficiency.

Discussion

In this study, we tested the effectiveness of Tekster, a comprehensive, strategy-focused writing instruction program developed for Dutch students in Grades 4–6. Participating teachers imple-

Table 5
Fit and Comparison of Nested Models

Model	$N_{\text{parameters}}$	−2LL	Comparison Models	 ΔX^2	 Δdf	 p
1 null	5	88763.76				
2 + measurement occasion (fixed)	7	88484.15	2vs1	279.61	2	<.001
3 + measurement occasion (random)	19	87826.54	3vs2	657.61	12	<.001
4 + group	20	87825.22	4vs3	1.32	1	.25
5 + intervention	21	87800.24	5vs4	24.98	1	<.001
6 + Intervention × Group	22	87800.12	6vs5	.12	1	.73

Table 6
Students' Average Writing Scores and Variances on Pre- and Posttest Measures

Measure	Measurement occasion		
	1	2	3
Fixed part			
Group 1	95.63 (1.38)	100.36 (1.36)	99.36 (1.34)
Group 2	98.54 (1.41)	98.78 (1.33)	103.51 (1.28)
Random part			
$S^2_{classes}$	53.92 (11.31)	49.79 (10.44)	43.73 (9.40)
S^2_{tasks}	9.20 (1.42)	9.20 (1.42)	9.20 (1.42)
$S^2_{students}$	59.99 (4.33)	54.98 (3.65)	54.31 (3.65)
S^2_{error}	128.48 (3.68)	99.11 (2.83)	92.98 (2.77)

Note. Standard errors are included in parentheses.

mented the intervention in their general education classrooms for a period of 8 weeks. Results indicated Tekster produced statistically significant improvements in the quality of students' texts. Students' individual writing quality did not only increase, but also became more consistent over time. The switching replication design allowed us to replicate the effect of the intervention within this study, as our findings demonstrate that the intervention was equally effective in both groups. Moreover, we found that students in Group 1 still wrote qualitatively better texts at the delayed posttest measure than at the pretest measure, indicating that the effect of the intervention was maintained after two months. Although there was a significant improvement of students' writing scores in all grades, the effect of the intervention was slightly smaller in Grade 6 than in Grades 4 and 5. Furthermore, results show that girls outperformed boys on all measurement occasions, but that the effect of the intervention was the same. Lastly, results of an aptitude treatment analysis showed that the effect of the intervention did not depend on students' writing proficiency.

Although the intervention was effective, the ES of the intervention on students' writing was moderate, 0.32. However, this ES is based on the average of completed lessons (which was 10) and is therefore a conservative estimate of the actual effect. Results showed the ES increased from 0.32 to 0.40 for students who completed all 16 lessons. Hence, students will make more progress if they complete the whole program. This can be achieved more easily when the implementation of Tekster is spread out over a longer period of time (e.g., one lesson a week), and/or if the program contained more lessons. Further research is needed to gain more insight in this aspect.

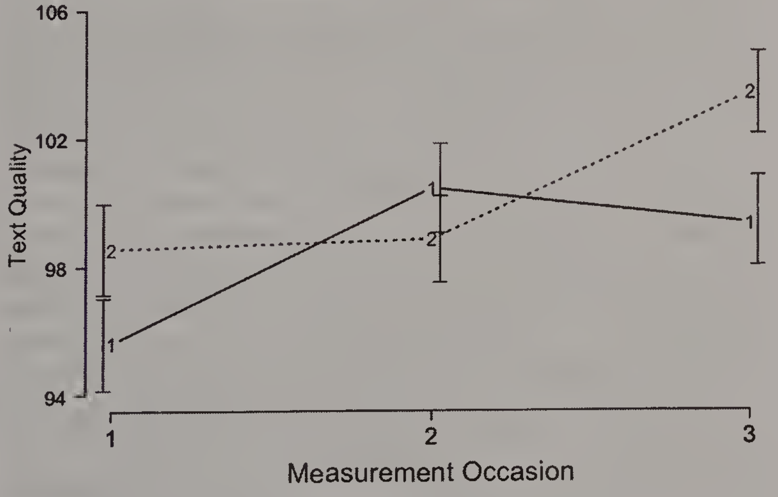


Figure 1. The effect of Tekster, averaged across all three grade levels. Error bars indicate 95% confidence intervals for the means. Solid lines represent Group 1, which received the intervention between first and second measurement occasion. Dashed lines represent Group 2, which received the intervention between second and third measurement occasion.

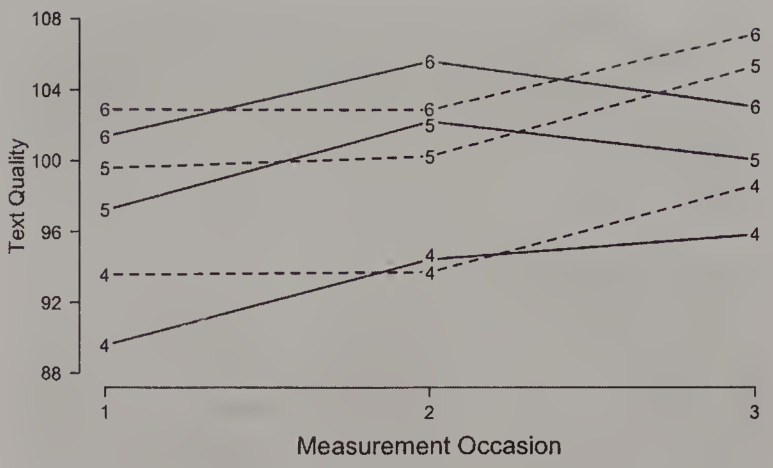


Figure 2. The effect of Tekster, by grade level. Solid lines represent Group 1, which received the intervention between first and second measurement occasion. Dashed lines represent Group 2, which received the intervention between second and third measurement occasion. Grade level is designated by the number (4, 5, or 6) shown with each line.

The effect of the intervention can also be interpreted in a more intuitive way by comparing it to the general improvement in writing skills of students between Grade 4 to 6 (Lipsey et al., 2012). Working with Tekster for 2 months resulted in an average gain in writing quality of 4.73 points. The average improvement in text quality scores between grades was 8.07 points, which means that students' writing improved by more than half a grade level.

Although Tekster was generally effective in improving students' writing performance, results showed that students' writing quality in Grade 4 and 5 improved slightly more than the writing quality of sixth grade students. An explanation for this can be that, even though the general approach is the same across grades, the acronyms differ slightly. Grade 6 is the only grade in which students are explicitly instructed to evaluate and revise. Research has shown that revising is difficult for students (Fitzgerald, 1987). To be able to revise, students must be aware of the goals and audience of their texts. In addition, they have to be able to critically read and evaluate their texts, and they have to know how they can fix problems, both on local and textual levels. Ideally, students start working with Tekster in Grade 4, when the focus is on learning and applying prewriting strategies, and gradually move on to Grade 6, when the focus shifts to revising. As this experiment was a cohort study, sixth graders lacked the basics that were the focus of instruction in Grade 4 and 5. We have addressed this issue by creating overlap in the topics that are covered in the different grades, but it might be that learning this overall approach at once was more complicated for Grade 6 students than the simpler versions of the acronym that were used in Grade 4 and 5. A longitudinal study would provide more insight in this matter.

A longitudinal study of Tekster would also shed more light on the learning trajectory of students across grades. The Dutch Inspectorate of Education (Henkens, 2010) reported that at present students hardly progress in their writing from Grade 4 to 6. As we have developed a systematic approach for the teaching of writing in the upper primary grades, we would expect a more continuous development of students' writing performance across the grades as a result.

Generalizability of the Results

In comparison to similar strategy-focused intervention studies aimed at Grade 4 to 6 in a general educational setting, the ES of this study (0.32) is notably smaller (cf. Graham et al., 2012; Koster et al., 2015, average ES 1.02 and 0.96, respectively). However, in contrast to most other intervention studies, Tekster was tested on a very large scale involving 1,420 students from 60 classes from 27 schools. Moreover, whereas most intervention studies used only one task as an indication of the effectiveness of their writing program, we tested students' overall writing proficiency with nine writing tasks in three genres: narrative, persuasive, and descriptive. Effects are therefore not only generalizable across students, but also across teachers and tasks. If we were to ignore the variance component related to tasks and classes, the ES of our intervention would increase to 0.63 and to 0.80 if the full program would have been completed, which is more in line with the effects reported in other intervention studies.

Maintenance Effects

Our results show that students' writing quality is still significantly above pretest level two months after the end of the program, which suggests that the intervention induced a lasting change in students' writing. We also see that students' writing scores did not continue to gain after the end of the intervention period. This is a mere illustration of Henkens's observation (Henkens, 2010) that the regular writing lessons in the average Dutch classroom do not lead to any significant improvement in students' writing. This is demonstrated in the present study by the fact that students in the control group (i.e., Group 2 between the first and the second measurement occasion) did not show any gains in writing quality.

It should be noted, however, that conclusions about the maintenance effect of the intervention are true only under the assumption that tasks were equally difficult and the effect of the intervention (i.e., interaction between condition and time) was the same for students in both conditions. Naturally, we tried to keep the writing tasks as similar as possible over the three measurement occasions, using the same rating procedure in which raters used the same benchmark scale for equal tasks across occasions, and calculating average scores based on three writing tasks per occasion. Nonetheless, we cannot entirely rule out the possibility that differences or similarities between scores over time (within conditions) are due to coincidence.

Effectiveness of Tekster for Different Types of Students

Results did not show an aptitude treatment interaction, indicating that all students, less proficient as well as proficient writers throughout Grade 4 to 6, benefited from the program to the same extent. This suggests that the program addressed the needs of all students, which is promising, given that in a general education classroom students differ considerably in their needs and abilities (Harris et al., 2012). The effectiveness of the program for different types of students can be explained in at least three ways. First, Tekster aimed to reduce cognitive overload during writing by providing students with skills and knowledge to regulate their writing process. Second, the program addressed the double challenge of writing and learning to write at the same time. Third, through Tekster's multifaceted approach, all students, weak as well as proficient writers, were provided with ample learning opportunities, for example by including coping as well as mastery peer modeling (Braaksma, 2002). That Tekster enhances the performance of all students is promising for whole classroom use, as a typical upper elementary classroom will contain students of various abilities.

Tekster's Effective Components

It should be noted that, although the program as a whole improved students' writing performance, we cannot make claims about the effectiveness of its individual components. We simply do not know which component is the most powerful ingredient of our approach. What we do know from previous research is that the combination of strategy-focused instruction and observational learning is highly effective in improving students' writing performance (Fidalgo et al., 2015). Fidalgo and colleagues assessed the effectiveness of four different instructional components of a strategy-focused writing training: modeling and reflection, direct instruction, peer feedback, and individual practice for sixth grade students, by manipulating the instructional sequence. Their results indicated that all positive effects are predominantly related to the modeling and reflection component. The way our study was designed does not allow for any conclusions regarding the effect of the observational learning component, but based on Fidalgo et al.'s (2015) findings, we suspect that, especially in combination with strategy-focused instruction, modeling may have contributed substantially to the effectiveness of our program. However, additional research is needed to isolate the influence of each component.

Teachers' Implementation of Tekster

Tekster was implemented by fourth, fifth, and sixth grade teachers in their own general education classrooms. Teachers from a large variety of schools participated in the study. Although this contributed considerably to the ecological validity of this study, it increased differences between classes. Furthermore, differences between teachers can also be caused by differences in teaching experience, background, teaching styles and individual preferences (Hattie, 2009). Hence, it is important to verify how teachers actually implemented the program in their classrooms. In previous studies, researchers often controlled for the differences between teachers by implementing the intervention themselves (e.g., Gordon & Braun, 1986; Kellogg, 1988) or by training teachers or teacher assistants intensively to implement the intervention (e.g., Fidalgo et al., 2015; Graham et al., 2005). Whereas intensive training is possible in a relatively small-scale study of one or two classes, this is not a feasible option when an intervention is implemented on a large scale.

The differences between classes can partly be explained by differences in the number of taught lessons. On average, 10 lessons were taught, but this number varied between classes, and we found that students' writing performance was positively related to the number of lessons taught. Furthermore, the results also showed that differences between teachers were reduced after the intervention, which suggests that teachers have adapted their instructional practice as a result of participation in the program. This seems to be confirmed by the fidelity measures, which revealed that teachers closely adhered to the lesson plans as indicated in the manual, and that they applied the key components of the intervention program, that is, modeling, the acronym, and the steps of the strategy.

It is promising that teachers were already capable of applying the key components of the program in their instruction after only a limited amount of training. However, the observational data do not allow for statements on the quality of the lessons, as they only provide information on what was done during the lessons. In further research, it is necessary to observe not only what teachers do in class, but also how they do this, for instance by videotaping and subsequently analyzing lessons to get a clearer picture of teachers' practices and whether and how they adapted the program to their own practice.

General Conclusion

To conclude, this study has shown that an overall approach in which several research-based instructional practices for teaching writing are combined is effective in improving elementary students' writing quality. This study is unique for the following reasons. First, through a switching replication design we were able to replicate the effect within one study, with the same results. Hence, the effects of the intervention do not seem to depend on characteristics of the sample. Together with the scale of the study, which included a large sample of Dutch schools, this allows us to make robust claims about the effectiveness of Tekster. Second, in this study we examined the impact of Tekster in a naturalistic setting, as the intervention was delivered in 60 general education classrooms by regular teachers, who were only trained for a short period of time. Third, students were taught a general strategy for writing, irrespective of genre, and the quality of their writing was measured with multiple writing tasks using multiple text types. It is therefore possible to generalize the results to overall writing proficiency in a general educational setting. All in all, this study demonstrates that a comprehensive writing program, such as Tekster, is a promising approach to improve elementary students' writing.

References

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.

Bean, T. W., & Steenwyk, F. L. (1984). The effect of three forms of summarization instruction on sixth graders' summary writing and composition. *Journal of Reading Behavior, 16,* 297–306.

Bereiter, C., Burtis, P. J., & Scardamalia, M. (1988). Cognitive operations in constructing main points in written composition. *Journal of Memory and Language, 27,* 261–278. http://dx.doi.org/10.1016/0749-596X(88)90054-X

Berninger, V., Yates, C., Cartwright, A., Rutberg, J., Remy, E., & Abbott, R. (1992). Lower-level developmental skills in beginning writing. *Reading and Writing, 4,* 257–280. http://dx.doi.org/10.1007/BF01027151

Blok, H., & Hoeksma, J. B. (1984). *Opstellen geschaald: De constructie van beoordelingsschalen voor vijf schrijfopdrachten* [Scaling essays: The construction of rating scales for five writing tasks]. Amsterdam, the Netherlands: Kohnstamm Institute.

Bodrova, E., & Leong, D. J. (1998). Scaffolding emergent writing in the zone of proximal development. *Literacy Teaching and Learning, 3,* 1–18.

Bouwer, R., Béguin, A., Sanders, T., & Van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing, 32,* 83–100. http://dx.doi.org/10.1177/0265532214542994

Bouwer, R., Koster, M., & Van den Bergh, H. (2016). *Benchmark rating procedure, best of both worlds? Comparing procedures to rate text quality in a reliable and valid manner*. Manuscript submitted for publication.

Braaksma, M. A. H. (2002). *Observational learning in argumentative writing* (Unpublished doctoral dissertation). University of Amsterdam, the Netherlands.

Braaksma, M. A. H., Rijlaarsdam, G., Van den Bergh, H., & Van Hout-Wolters, B. H. A. M. (2004). Observational learning and its effect on the orchestration of writing processes. *Cognition and Instruction, 22,* 1–36. http://dx.doi.org/10.1207/s1532690Xci2201_1

Brunstein, J. C., & Glaser, C. (2011). Testing a path-analytic mediation model of how self-regulated writing strategies improve fourth graders' composition skills: A randomized controlled trial. *Journal of Educational Psychology, 103,* 922–938. http://dx.doi.org/10.1037/a0024622

Central Office for Statistics. (2015, July 15). *(Speciaal) basisonderwijs; culturele minderheden, (achterstands)leerlingen* [(Special) primary education; cultural minority groups, (disadvantaged)students]. Retrieved from http://statline.cbs.nl/StatWeb/publication/?VW=T&DM=SLNL&PA=37846SOL&D1=0&D2=a&D3=a&D4=a&HD=090218-1354&HDR=T,G2,G1&STB=G3

Couzijn, M. J. (1995). *Observation of writing and reading activities: Effects on learning and transfer* (Unpublished doctoral dissertation). University of Amsterdam, the Netherlands.

Couzijn, M. (1999). Learning to write by observation of writing and reading processes: Effects on learning and transfer. *Learning and Instruction, 9,* 109–142. http://dx.doi.org/10.1016/S0959-4752(98)00040-1

Couzijn, M., & Rijlaarsdam, G. (2004). Learning to write by reader observation and written feedback. In G. Rijlaarsdam, H. van den Bergh, & M. Couzijn (Eds.), *Effective teaching and learning of writing. Current trends in research* (pp. 224–252). Amsterdam, the Netherlands: Amsterdam University Press.

Crowhurst, M. (1990). Reading/writing relationships: An intervention study. *Canadian Journal of Education/Revue canadienne de l'éducation, 15,* 155–172. http://dx.doi.org/10.2307/1495373

Crowhurst, M. (1991). Interrelationships between reading and writing persuasive discourse. *Research in the Teaching of English, 25,* 314–338.

Department for Education. (2012). *What is the research evidence on writing?* (Research Report No. DFE-RR238). Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/183399/DFE-RR238.pdf

Expert Group Learning Trajectories. (2009). *Referentiekader taal en rekenen: De referentieniveaus* [Reference framework language and arithmetic: The referential levels]. Retrieved from http://www.taalenrekenen.nl/downloads/referentiekader-taal-en-rekenen-referentieniveaus.pdf

Fayol, M. (1999). From on-line management problems to strategies in written composition. In M. Torrance & G. Jeffery (Eds.), *The cognitive demands of writing: Processing capacity and working memory effect in text production* (pp. 13–23). Amsterdam, the Netherlands: Amsterdam University Press.

Ferretti, R. P., Lewis, W. E., & Andrews-Weckerly, S. (2009). Do goals affect the structure of students' argumentative writing strategies? *Journal of Educational Psychology, 101,* 577–589. http://dx.doi.org/10.1037/a0014702

Ferretti, R. P., MacArthur, C. A., & Dowdy, N. C. (2000). The effects of an elaborated goal on the persuasive writing of students with learning disabilities and their normally achieving peers. *Journal of Educational Psychology, 92,* 694–702. http://dx.doi.org/10.1037/0022-0663.92.4.694

Fidalgo, R., Torrance, M., Rijlaarsdam, G., Van den Bergh, H., & Lourdes Álvarez, M. (2015). Strategy-focused writing instruction: Just observing and reflecting on a model benefits sixth grade students. *Contemporary Educational Psychology, 41,* 37–50. http://dx.doi.org/10.1016/j.cedpsych.2014.11.004

Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research, 57,* 481–506. http://dx.doi.org/10.3102/00346543057004481

Fitzgerald, J., & Markham, L. R. (1987). Teaching children about revision in writing. *Cognition and Instruction, 4,* 3–24. http://dx.doi.org/10.1207/s1532690xci0401_1

Fitzgerald, J., & Teasley, A. B. (1986). Effects of instruction in narrative structure on children's writing. *Journal of Educational Psychology, 78,* 424–432. http://dx.doi.org/10.1037/0022-0663.78.6.424

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32,* 365–387. http://dx.doi.org/10.2307/356600

Gordon. C. J., & Braun. C. (1986). Mental processes in reading and writing: A critical look at self-reports as supportive data. *The Journal of Educational Research, 79,* 292–301. http://dx.doi.org/10.1080/00220671.1986.10885694

Graham, S. (2006). Strategy instruction and the teaching of writing: A meta-analysis. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 187–207). New York, NY: Guilford Press.

Graham, S., Harris, K., & Hebert, M. (2011). It is more than just the message: Analysis of presentation effects in scoring writing. *Focus on Exceptional Children, 44,* 1–12.

Graham, S., Harris, K., & Mason, L. (2005). Improving the writing performance, knowledge, and self-efficacy of struggling young writers: The effects of self-regulated strategy development. *Contemporary Educational Psychology, 30,* 207–241. http://dx.doi.org/10.1016/j.cedpsych.2004.08.001

Graham, S., MacArthur, C., & Schwartz, S. (1995). Effects of goal setting and procedural facilitation on the revising behavior and writing performance of students with writing and learning problems. *Journal of Educational Psychology, 87,* 230–240. http://dx.doi.org/10.1037/0022-0663.87.2.230

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology, 104,* 879–896. http://dx.doi.org/10.1037/a0029185

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99,* 445–476. http://dx.doi.org/10.1037/0022-0663.99.3.445

Harris, K. R., Graham, S., Mason, L. H., & Saddler, B. (2002). Developing self-regulated writers. *Theory into Practice, 41,* 110–115. http://dx.doi.org/10.1207/s15430421tip4102_7

Harris, K. R., Lane, K. L., Graham, S., Driscoll, S. A., Sandmel, K., Brindle, M., & Schatschneider, C. (2012). Practice-based professional development for self-regulated strategies development in writing a randomized controlled study. *Journal of Teacher Education, 63,* 103–119. http://dx.doi.org/10.1177/0022487111429005

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London, UK: Routledge.

Henkens, L. S. J. M. (2010). *Het onderwijs in het schrijven van teksten* [Education in text writing]. Utrecht, the Netherlands: Inspectorate of Education.

Hillocks, G., Jr. (1984). What works in teaching composition: A meta-analysis of experimental treatment studies. *American Journal of Education, 93,* 133–170. http://dx.doi.org/10.1086/443789

Hintze, J. M., Volpe, R. J., & Shapiro, E. S. (2002). Best practices in systematic direct observation of student behavior. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 993–1006). Washington, DC: National Association of School Psychologists.

Holliway, D. R., & McCutchen, D. (2004). Audience perspective in young writers' composing and revising: Reading as the reader. In G. Rijlaarsdam (Series Ed.) & L. Allal, P. Chanquoy, & P. Largy (Vol. Eds.), *Studies in writing: Vol. 13. Revision: Cognitive and instructional processes* (pp. 105–121). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Inspectorate of Education. (2012). *De staat van het onderwijs. Onderwijsverslag 2010/2011* [The state of education. Educational report 2010/2011]. Retrieved from http://www.onderwijsinspectie.nl/binaries/content/assets/Onderwijsverslagen/2012/onderwijsverslag_2010_2011_printversie.pdf

Kellogg, R. T. (1988). Attentional overload and writing performance: Effects of rough draft and outline strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 355–365.

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research, 1,* 1–26. http://dx.doi.org/10.17239/jowr-2008.01.01.1

Koopman, P., Ledoux, G., Karssen, M., Van der Meijden, A., & Petit, R. (2015). *Vervolgmeting 1 kengetallen passend onderwijs* [Sequel measure 1 key ratios inclusive education] (Rapport 936, No. 20667). Amsterdam, the Netherlands: Kohnstamm Institute.

Koster, M., Bouwer, R., & Van den Bergh, H. (2014a). *VOS: Werkboek en docentenhandleiding voor groep 6* [FOX: Workbook and teacher manual for grade 4]. Utrecht, the Netherlands: Utrecht University.

Koster, M., Bouwer, R., & Van den Bergh, H. (2014b). *DODO: Werkboek en docentenhandleiding voor groep 7* [DODO: Workbook and teacher manual for grade 5]. Utrecht, the Netherlands: Utrecht University.

Koster, M., Bouwer, R., & Van den Bergh, H. (2014c). *EKSTER: Werkboek en docentenhandleiding voor groep 8* [MAGPIE: Workbook and teacher manual for grade 6]. Utrecht, the Netherlands: Utrecht University.

Koster, M., Bouwer, R., & Van den Bergh, H. (2016). *A letter of advice to Like: Examining the relationship between metacognitive knowledge and writing performance.* Manuscript submitted for publication.

Koster, M., Tribushinina, E., De Jong, P. F., & Van den Bergh, H. (2015). Teaching children to write: A meta-analysis of writing intervention research. *Journal of Writing Research, 7,* 249–274. http://dx.doi.org/10.17239/jowr-2015.07.02.2

Kuhlemeier, H., Til, A. V., Hemker, B., de Klijn, W., & Feenstra, H. (2013). *Balans van de schrijfvaardigheid in het basis- en speciaal basisonderwijs 2* [Present state of writing competency in elementary and special education 2] (PPON Report No. 53). Arnhem, the Netherlands: Cito.

Latham, G. P., & Locke, E. A. (1991). Self-regulation through goal setting. *Organizational Behavior and Human Decision Processes, 50,* 212–247. http://dx.doi.org/10.1016/0749-5978(91)90021-K

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (NCSER 2013–3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review, 8,* 299–325. http://dx.doi.org/10.1007/BF01464076

McCutchen, D. (2011). From novice to expert: Implications of language skills and writing-relevant knowledge for memory during the development of writing skill. *Journal of Writing Research, 3,* 51–68. http://dx.doi.org/10.17239/jowr-2011.03.01.3

Midgette, E., Haria, P., & MacArthur, C. (2008). The effects of content and audience awareness goals for revision on the persuasive essays of fifth- and eighth-grade students. *Reading and Writing, 21,* 131–151. http://dx.doi.org/10.1007/s11145-007-9067-9

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97. http://dx.doi.org/10.1037/h0043158

Ministry of Education, Culture, and Science. (2015). *Basisonderwijs 2015–2016* [Primary Education 2015–2016]. Den Haag, the Netherlands: Author.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. http://dx.doi.org/10.1126/science.aac4716

O'Sullivan, J., & Pressley, M. (1984). Completeness of instruction and strategy transfer. *Journal of Experimental Child Psychology, 38,* 275–288. http://dx.doi.org/10.1016/0022-0965(84)90126-7

Parr, J. M., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing Writing, 15,* 68–85. http://dx.doi.org/10.1016/j.asw.2010.05.004

Pearson, P. D., & Gallagher, G. (1983). The instruction of reading comprehension. *Contemporary Educational Psychology, 8,* 317–344. http://dx.doi.org/10.1016/0361-476X(83)90019-X

Peterson, S. (2000). Grades four and eight students' and teachers' perceptions of girls' and boys' writing competencies. *Reading Horizons, 40,* 253–271.

Raedts, M., Rijlaarsdam, G., Van Waes, L., & Daems, F. (2007). Observational learning through video-based models: Impact on students' accuracy of self-efficacy beliefs, task knowledge and writing performances. In G. Rijlaarsdam, P. Boscolo, & S. Hidi (Eds.), *Studies in writing: Writing and motivation* (Vol. 19, pp. 219–238). Oxford, UK: Elsevier.

Raphael, T. E., & Kirschner, B. M. (1985). *The effects of instruction in compare/contrast text structure on sixth-grade students' reading comprehension and writing products* (Research Series No. 161). East Lansing, MI: Michigan State University, Institute for Research on Teaching.

Rietdijk, S., Van Weijen, D., Janssen, T., Van den Bergh, H., & Rijlaarsdam, G. (2015). *Teaching writing in primary education: Classroom practices, learning time, and teacher characteristics and their relationships.* Manuscript submitted for publication.

Rijlaarsdam, G. (2005). Observerend leren: Een kernactiviteit in taalvaardigheidsonderwijs [Observational learning: A core activity in language education]. *Levende Talen Tijdschrift, 6,* 10–28.

Rijlaarsdam, G., Braaksma, M., Couzijn, M., Janssen, T., Raedts, M., Van Steendam, E., . . . Van den Bergh, H. (2008). Observation of peers in learning to write. practice and research. *Journal of Writing Research, 1,* 53–83. http://dx.doi.org/10.17239/jowr-2008.01.01.3

Rijlaarsdam, G., & Couzijn, M. (2000). Writing and learning to write: A double challenge. In R. Simons, J. van der Linden, & T. Duffy (Eds.), *New learning* (pp. 157–189). Dordrecht, the Netherlands: Kluwer Academic Publishers. http://dx.doi.org/10.1007/0-306-47614-2_9

Rijlaarsdam, G., Couzijn, M., Janssen, T., Braaksma, M., & Kieft, M. (2006). Writing experiment manuals in science education: The impact of writing, genre, and audience. *International Journal of Science Education, 28,* 203–233. http://dx.doi.org/10.1080/09500690500336932

Rijlaarsdam, G., Janssen, T., Rietdijk, S., & Van Weijen, D. (in press). Reporting design principles for effective instruction of writing: Intervention as constructs. In R. Fidalgo, K. Harris, & M. Braaksma (Eds.), *Design principles for teaching effective writing: Theoretical and empirical grounded principles.* Leiden, the Netherlands: Brill Publishers.

Salahu-Din, D., Persky, H., & Miller, J. (2008). *The nation's report card: Writing 2007* (NCES 2008–468). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Scardamalia, M., & Paris, P. (1985). The function of explicit discourse knowledge in the development of text representations and composing strategies. *Cognition and Instruction, 2,* 1–39. http://dx.doi.org/10.1207/s1532690xci0201_1

Schoonen, R., & De Glopper, K. (1996). Writing performance and knowledge about writing. In G. Rijlaarsdam, H. van den Bergh, & M. Couzijn (Eds.), *Theories, models, and methodology in writing research* (pp. 87–107). Amsterdam, the Netherlands: Amsterdam University Press.

Schriver, K. A. (1992). Teaching writers to anticipate reader's needs: A classroom pedagogy. *Written Communication, 9,* 179–208. http://dx.doi.org/10.1177/0741088392009002001

Schunk, D. H. (1987). Peer models and children's behavioral change. *Review of Educational Research, 57,* 149–174. http://dx.doi.org/10.3102/00346543057002149

Schunk, D. H. (1990). Goal setting and self-efficacy during self-regulated learning. *Educational Psychologist, 25,* 71–86. http://dx.doi.org/10.1207/s15326985ep2501_6

Schunk, D. H. (2012). Social cognitive theory. In D. Schunk (Ed.), *Learning theories: An educational perspective* (6th ed., pp. 117–162). Boston, MA: Pearson.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin Company.

Solomon, S. R., & Sawilowsky, S. S. (2009). Impact of rank-based normalizing transformations on the accuracy of test scores. *Journal of Modern Applied Statistical Methods, 8,* 448–462.

Troia, G. A., & Graham, S. (2002). The effectiveness of a highly explicit, teacher-directed strategy instruction routine: Changing the writing performance of students with learning disabilities. *Journal of Learning Disabilities, 35,* 290–305. http://dx.doi.org/10.1177/00222194020350040101

Van den Bergh, H., & Eiting, M. H. (1989). A method of estimating rater reliability. *Journal of Educational Measurement, 26,* 29–40. http://dx.doi.org/10.1111/j.1745-3984.1989.tb00316.x

Van Steendam, E., Rijlaarsdam, G., Van den Bergh, H., & Sercu, L. (2014). The mediating effect of instruction on pair composition in L2 revision and writing. *Instructional Science, 42,* 905–927. http://dx.doi.org/10.1007/s11251-014-9318-5

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Wong, B. Y., Hoskyn, M., Jai, D., Ellis, P., & Watson, K. (2008). The comparative efficacy of two approaches to teaching sixth graders opinion essay writing. *Contemporary Educational Psychology, 33,* 757–784. http://dx.doi.org/10.1016/j.cedpsych.2007.12.004

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 17,* 89–100. http://dx.doi.org/10.1111/j.1469-7610.1976.tb00381.x

Zimmerman, B. J., & Kitsantas, A. (2002). Acquiring writing revision and self-regulatory skill through observation and emulation. *Journal of Educational Psychology, 94,* 660–668. http://dx.doi.org/10.1037/0022-0663.94.4.660

Zimmerman, B. J., & Risemberg, R. (1997). Becoming a self-regulated writer: A social cognitive perspective. *Contemporary Educational Psychology, 22,* 73–101. http://dx.doi.org/10.1006/ceps.1997.0919

Received May 13, 2015
Revision received March 19, 2017
Accepted March 20, 2017 ■

Journal of Educational Psychology
2018, Vol. 110, No. 1, 72–86

0022-0663/18/$12.00 http://dx.doi.org/10.1037/edu0000198

Understanding How Syntactic Awareness Contributes to Reading Comprehension: Evidence From Mediation and Longitudinal Models

S. Hélène Deacon
Dalhousie University

Michael Kieffer
New York University

The authors tested theoretically driven predictions as to the ways in which syntactic awareness, or awareness of word order within sentences, might contribute to reading comprehension, the end goal of reading development and instruction. They conducted a longitudinal study of 100 English-speaking children followed from Grade 3 to 4. Children completed measures of syntactic awareness, word reading, reading comprehension, and reading-related control variables. Path analyses at each of Grades 3 and 4 show a unique concurrent relation of syntactic awareness with reading comprehension, but not to word reading skills. Longitudinal analyses reveal that syntactic awareness at Grade 3 predicts gains in reading comprehension between Grades 3 and 4. Together, findings suggest a robust role for syntactic awareness in the development of reading comprehension.

Educational Impact and Implications Statement

Children's ability to understand texts is essential to success in education and beyond. Through elementary school, children's texts become increasingly complex, particularly in their sentence structure. The authors show that children's awareness of sentence structure, or their syntactic awareness, determines the gains that they make in their reading comprehension between Grades 3 and 4. These findings suggest that supporting children in developing syntactic awareness is likely to have knock-on effects to their understanding of complex texts.

Keywords: syntactic awareness, reading comprehension

Reading has been widely described as fundamentally metalinguistic; it involves the ability to reflect on and manipulate the structural features of language (e.g., Nagy, 2007; Tumner, Herriman, & Nesdale, 1988). Awareness of some of these structural features has received more empirical attention than others. For example, a vast body of research demonstrates the importance of phonological awareness in word reading (for reviews, see National Institute of Child Health and Human Development, 2000; Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001; see also Bradley & Bryant, 1983; Perfetti, Beck, Bell, & Hughes, 1987). Far less research has targeted syntactic awareness, or skill in "reflecting on or manipulating the order of words in a sentence" (Nagy, 2007, p. 53), which might be particularly important for reading comprehension (e.g., Demont & Gombert, 1996; Muter, Hulme, Snowling, & Stevenson, 2004). Here, we test theoretically predicted relations between syntactic awareness and reading comprehension.

Syntactic awareness is often measured with word order correction or judgment tasks. For example, in the correction task in Tumner et al.'s (1988) study, children were asked to rearrange three to four orally presented words into sentences, such as "dog the barked." In a judgment task, Geva and Farnia (2012) asked children to judge the correctness of both correct and incorrect sentences (e.g., "Has the king been served his dinner?" and "Has been the king served his dinner?", respectively). These tasks tap syntax in their emphasis on word order. As such, these tasks are suggested to capture individual differences in children's ability to reflect on or manipulate word order at the sentence-level. As we will see, however, many tasks in studies completed to date have also incorporated other forms of linguistic information, such as morphology. Morphology is considered to be a separate linguistic, and therefore metalinguistic, construct from syntax (e.g., Bowey, 1994; Crystal, 1987; Kuo & Anderson, 2006). As such, the current study includes tasks designed to specifically measure children's awareness of syntax.

Classic and current theories of reading comprehension broadly agree that syntactic awareness plays a direct role in reading comprehension (Perfetti & Stafura, 2014; RAND Reading Study Group, 2002; Scarborough, 2002; see horizontal arrow in Figures 1a and 1b). The theoretical rationale is that syntactic awareness allows children to break complex sentences into more manageable chunks, facilitating their formation of a representation for the whole text (e.g., Verhoeven & Perfetti, 2008). Syntactic awareness

This article was published Online First April 13, 2017.

S. Hélène Deacon, Department of Psychology and Neuroscience, Dalhousie University; Michael Kieffer, Department of Teaching and Learning, New York University.

Correspondence concerning this article should be addressed to S. Hélène Deacon, Department of Psychology and Neuroscience, Dalhousie University, 1355 Oxford Street Halifax, Nova Scotia, Canada B3H 4R2. E-mail: helene.deacon@dal.ca

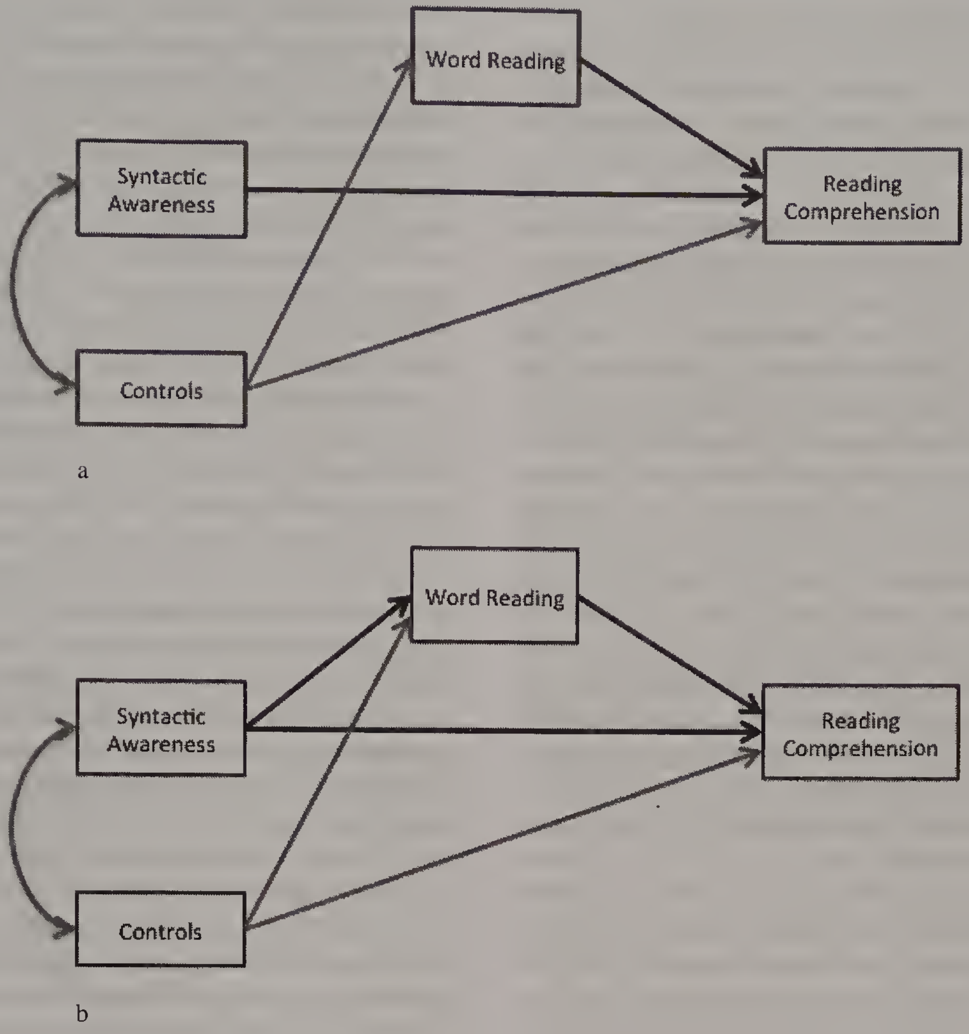


Figure 1. Contrasting hypothesized models for the relations of syntactic awareness to reading comprehension as independent of word reading (a) and as partially mediated by word reading (b). a: Predictions of direct relations between syntactic awareness and reading comprehension, with no mediation by word reading. b: Predictions of direct relations between syntactic awareness and reading comprehension, with mediation by word reading.

is particularly relevant in this process because it reflects the child's awareness of sentence structures in general. Take an example from the Grade 3 Ontario provincial reading assessment: "I kept my eye on my swimming tree, the tall pine I always kept in sight, ever since I was little and first learned to swim across the lake" (Education Quality and Accountability Office, 2015). Breaking this sentence into its 4 chunks (or clauses) would be very useful in figuring out its meaning in comparison to trying to simultaneously process its 29 words as one larger chunk. Building on this, syntactic awareness is often conceptualized as supporting the parsing of sentences into smaller components, which are in turn recombined into text level representations (see also e.g., Farnia & Geva, 2013; Willows & Ryan, 1986). This is the basis for the prediction that awareness of sentence structure directly supports reading comprehension.

The key theoretical debate lies in whether syntactic awareness also supports reading comprehension indirectly, specifically via word reading. Several approaches predict an indirect relation (see Figure 1b). As an example, Perfetti, Landi, and Oakhill (2005) specify two roles for syntax. The first is as a part of higher-order language comprehension processes, supporting a direct role. The second emerges from the suggestion that syntax is a part of high quality lexical representations (Perfetti & Hart, 2002), supporting an indirect role for syntactic awareness in reading comprehension through word reading (see also Perfetti & Stafura, 2014). Tumner (1989) made the same prediction of indirect relations, but articulated an alternative mechanism; children could use their awareness of sentence structures to combine partial decoding with contextual cues to arrive at the correct word pronunciation. Regardless of these differences in mechanism, both approaches predict a direct role of syntactic awareness in reading comprehension, as well as a mediated one through word reading (see Figure 1b).

In contrast, multiple models of word reading predict no role for syntactic awareness in word reading (e.g., Ehri, 2005, 2015; Perfetti, 2007; Share, 2008; see Figure 1a); as such, there is no predicted indirect relation through word reading to reading comprehension. These models are quite clear that reliance on syntactic information should have no role in word reading, unless it is a negative one (e.g., Ehri, 2005). Instead, they emphasize the importance of automatic decoding of each word within a text, with no reliance on sentence structure to support this (see also Rayner, 1998).

Here we test the competing predictions generated by these models. Specifically, we test whether there are indirect associations of syntactic awareness on reading comprehension via word reading or not. Answering this question requires mediation modeling, which has been applied relatively rarely in this area of research.

A good deal of prior research has demonstrated a relation between syntactic awareness and word reading and/or reading comprehension in children across the elementary school years (e.g., Blackmore & Pratt, 1997; Browne Rego & Bryant, 1993; Holsgrove & Garton, 2006; Klauda & Guthrie, 2008; but see, e.g., Cain, 2007). These studies have typically included a strong set of controls, such as phonological awareness (Holsgrove & Garton, 2006) and nonverbal ability, vocabulary, and verbal memory (Browne Rego & Bryant, 1993). A smaller set of studies show that the relation between syntactic awareness and reading comprehension survives controls for word reading (e.g., Bowey, 1986; Brimo, Apel, & Fountain, 2017; Geva & Farnia, 2012). As such, the relation between syntactic awareness and reading comprehension does not appear to be fully explainable via the relation between either one and word reading. This is a first step in establishing a direct relation between syntactic awareness and reading comprehension. Critically, though, these studies have not assessed possible mediation by word reading and, therefore, cannot speak to the possibility of the indirect relation.

The most direct test of the theoretically contentious indirect relation is via mediation modeling, an approach used in two studies to date. In each of these longitudinal studies, syntactic awareness was measured with word order correction tasks, which required children to rearrange a set of words into complete sentences. Tumner et al. (1988) reported on analyses of the role of syntactic awareness using data collected from children at the end of Grade 1. In modeling controlling for phonological awareness and pragmatic awareness, syntactic awareness predicted reading comprehension indirectly via nonword decoding. There was no significant direct relation to reading comprehension. Tumner (1989) reported on modeling with data from both Grades 1 and 2, with controls for phonological awareness, verbal ability, and concrete operativity. At Grades 1 and 2, syntactic awareness predicted reading comprehension indirectly via nonword decoding. At Grade 2, there was an additional indirect relation through listening comprehension. To our knowledge, Tumner (1989)'s models did not evaluate the possible direct relation to reading comprehension. Together, these studies provide good evidence of indirect relations from syntactic awareness to reading comprehension via nonword decoding in young readers; however, there have been few empirical tests of the predicted direct relations.

Moving forward, it would be important to evaluate these relations with more complete modeling of indirect and direct relations in a study of older children. In terms of modeling, it would be useful to include both direct and indirect relations in the model, contrasting model fit to further evaluate the nature of the relation between syntactic awareness and reading comprehension. In terms of age range, the high correlation between word reading and reading comprehension in young children (e.g., Gough, Hoover, & Peterson, 1996) makes it challenging to evaluate mediation in a meaningful way. As children progress through the elementary school years, there is a greater divergence between their word reading and reading comprehension skill (e.g., Gough et al., 1996). This is in part because texts become increasingly complex over the elementary school period, especially in their syntactic structures (e.g., Williamson, Fitzgerald, & Stenner, 2013). As such, it seems particularly relevant to evaluate these theoretically predicted relations with children who have progressed beyond the initial stages of learning (e.g., Gottardo, Stanovich, & Siegel, 1996; Muter & Snowling, 1998; Tumner et al., 1988).

Our second research question addresses the fact that the competing hypotheses tested here are derived from models of reading acquisition, whether these are of word reading or reading comprehension (e.g., Ehri, 2005; Perfetti et al., 2005). As such, any model that predicts a role for syntactic awareness in either word reading or reading comprehension effectively hypothesizes relations to gains in children's skill over time. Such relations are depicted in cross-path *a* in Figure 2 in relation to reading comprehension. Accordingly, a strong test of the predicted models lies in determining whether there is evidence for theoretically conceptualized variables as longitudinal predictors of later reading development. Scarborough (2002), for example, predicted that children better able to reflect on the phrase and sentence level of language should make more progress in learning to understand texts. As such, the prediction is that syntactic awareness is not only related to reading comprehension at a single point in time, but also to gains in

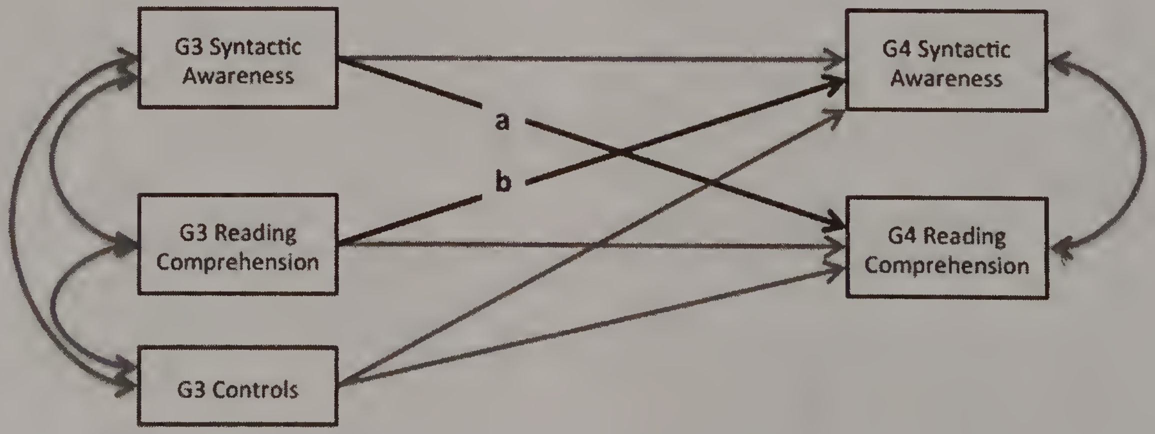


Figure 2. Hypothesized model evaluating the relation between Grade (G) 3 syntactic awareness and gains in reading comprehension (cross-path a) and the relation between Grade 3 reading comprehension and gains in syntactic awareness (cross-path b).

reading comprehension over time. Importantly, this longitudinal prediction about reading acquisition is the basis for all recommendations for practice that are derived from theoretical models.

In addition, it is important to consider that relations have also been predicted from reading comprehension to syntactic awareness (see cross-path *b* in Figure 2). Perfetti et al. (2005) suggested that successful reading offers "experience with syntactic structures that are less common in spoken than in written language" (2005, p. 238; see also Chafe, 1985; Ehri, 1976). This seems particularly plausible given the relatively greater complexity of syntactic structures in print than in oral language (e.g., Williamson et al., 2013). Further, seeing complex syntax in print offers children more opportunities to reflect on these structures through reading and rereading. As in the other direction, the key test of these theories lies in whether reading comprehension predicts gains in syntactic awareness over time. As such, the current study investigates both relations, which is our second research question.

A highly effective way to evaluate whether syntactic awareness contributes to gains in reading comprehension lies in the use of auto-regressor analyses of a longitudinal study (Kenny, 1975). As an example, one can assess whether early syntactic awareness predicts gains in reading comprehension. Controlling for the auto-regressor, or the prior level of the outcome variable, permits the conclusion that early syntactic awareness is associated with change in, and not merely later levels of, reading comprehension. These analyses can also evaluate whether early reading comprehension is associated with gains in syntactic awareness.

To our knowledge, two studies have evaluated the temporal order of the relation between syntactic awareness and reading comprehension using auto-regressor analyses. These extend the large body of research conducted using a cross-sectional or longitudinal design (e.g., Bowey, 1986; Willows & Ryan, 1986). The two studies that we focus on here are unique in that they have been conducted with a longitudinal design and included controls for prior levels of reading (Blackmore & Pratt, 1997; Farnia & Geva, 2013). In both studies, children were asked to either correct or judge the accuracy of sentences that included syntactic and morphological errors (e.g., "Jane saw two horse"). In terms of results, Blackmore and Pratt (1997) showed that syntactic awareness at 5 years did not predict reading comprehension at 6 years, beyond controls for word reading at age 5, vocabulary, and phonological awareness. Word reading was the auto-regressor, likely due to challenges in measuring reading comprehension in young children. With older children, Farnia and Geva (2013) found that syntactic awareness at each of Grades 1 and 4 predicted the rate of growth in reading comprehension between Grades 4 and 6, beyond word reading and other variables. Together, these studies suggest that syntactic awareness may predict progress in reading comprehension for older, but not necessarily younger, readers.

One needs to be cautious in these conclusions, however. By including both morphological and syntactic errors, measurement of the syntactic awareness construct in both studies introduced a potential confound with morphological awareness. This has been an issue in a good deal of prior work (e.g., Bowey, 1986; Geva & Farnia, 2012). This is of concern given that morphological awareness is known to be related to both word reading and reading comprehension (e.g., Deacon & Kirby, 2004). As such, it would be important to evaluate carefully the possibility that syntactic awareness remains a predictor of gains in reading comprehension when there is no influence of morphological awareness. Further, we need to assess whether early reading comprehension skill predicts gains in syntactic awareness, a possibility that has yet to be evaluated.

In the present study, we followed a group of children from Grade 3 to Grade 4 to investigate relations between syntactic awareness and both word reading and reading comprehension. This is a point in development at which word reading and reading comprehension clearly diverge (e.g., Chall, 1983). We first model the concurrent relations between syntactic awareness, word reading, and reading comprehension (following on Figure 1) with multivariate path analyses, including multiple controls. This allows us to identify whether syntactic awareness has a unique role in word reading and/or reading comprehension, as well as potential mediated relations. We then model longitudinal relations between syntactic awareness and gains in both word reading skills and reading comprehension. These analyses provide conservative, if correlational, evidence for the theoretically driven predictions of the relations between syntactic awareness and reading comprehension.

In this study, we designed our measure of syntactic awareness to isolate effects from other factors. Our task includes syntactic errors only, avoiding the morphological errors included in a good deal of prior work (e.g., Bowey, 1986) that have drawn criticism (Bowey, 1994; Kuo & Anderson, 2006). Our measurement of syntactic awareness also works to ensure that children have more than adequate linguistic knowledge to complete the awareness task. We chose syntactic structures that are produced accurately by 5 years of age, three years younger than the children in our study. Specifically, the syntactic structures in our task are mastered in naturalistic speech by 5 years of age (e.g., relative clauses, violation of SVO order; Brown, 1973), with confirmation of this pattern in in experimental studies (e.g., Childers & Tomasello, 2002; Chomsky, 1969; Kidd & Bavin, 2002; Pierce, 1992). Further, the items in our task have only one or two clauses; 8-year-old children routinely form sentences with multiple clauses far more complex than those included in our task (Gillam, Marquardt, & Martin, 2010). These design decisions were made to increase confidence that results are due to syntactic awareness per se rather than other factors.

Finally, we performed conservative tests of these theoretical predictions of a specific role for syntactic awareness in reading development by including multiple relevant control variables that account for theoretically viable counter hypotheses. In terms of nonverbal controls, we assessed nonverbal reasoning. This is a standard control in studies of reading development (e.g., Deacon & Kirby, 2004) in part because many language tasks, including syntactic awareness, draw, at least in part, on nonlinguistic reasoning skills. It was also important to assess other language skills, so that we could have confidence that any relations we uncovered were specific to syntactic awareness, and not general to language skills. In this study, we assessed vocabulary and morphological awareness, in large part because of their well-demonstrated roles in the development of reading comprehension in the age range studied here. Specifically, vocabulary, or knowledge of word meaning, is a powerful predictor of children's gains in reading comprehension over the age range studied here (Verhoeven & Van Leeuwe, 2008). Further, there is highly consistent evidence that morphological awareness, or awareness of the minimal units in language, predicts gains in reading comprehension in children of this age range (Deacon, Kieffer, & Laroche, 2014; Deacon & Kirby, 2004;

Foorman, Petscher, & Bishop, 2012; Kruk & Bergman, 2013). We also controlled for children's age to account for relations between reading outcomes and variation in age within grade level. In all analyses we compared the magnitude of predictive relations from syntactic awareness to those of other known predictors, enabling us to evaluate the relative importance of syntactic awareness.

Method

Participants

A total of 100 English first-language children participated in this longitudinal study across Grades 3 and 4. These children were from rural schools on the east coast of Canada. At Grade 3, the children were a mean age of 8 years 11 months (*SD* = 3 months), and there were 47 boys and 53 girls.

These children were originally recruited as a part of a larger longitudinal study conducted from Grades 1 to 4. All children in Grade 1 in the seven participating schools were invited to participate; only those with written parental consent and verbal child assent were tested. Average participation rate was 62%. A total of 124 children were originally recruited at Grade 1, with the majority of attrition due to family mobility. There were no significant differences in scores at Grade 1 between children who remained in the study through to Grade 4 and those who did not.

In a questionnaire sent home with the children, parents reported being working class on average (based on Hollingshead, 1957; *M* = 4.17, *SD* = 1.77). Mean standard scores indicate that the sample performed in the average range on all standardized measures administered (see Table 1). Most scores on the standardized tasks are just a few points above average and with a slightly smaller average standard deviation. As such, the sample can be considered relatively representative.

Procedure

All tasks were administered individually to the participants in a quiet room in the children's schools. All testers had at least an undergraduate degree in Psychology and were extensively trained prior to administering tasks. Tasks were presented in the same order to all participants. All measures, save one (nonverbal reasoning), were administered to participants in both Grades 3 and 4. Nonverbal reasoning was assessed only in Grade 3 due to its developmental stability (Wechsler, 1999). Alternate forms were used across years where available (i.e., for Passage Comprehension and Word Identification and Attack). There was approximately 12 months between testing points. The tasks took approximately 1.5 hr to complete. Tasks were administered in two to four shorter sessions depending on the child's interest level and classroom schedules. Tasks only had time limits when these were stipulated in the manual's instructions for standardized tasks.

Measures

For all measures, please see Table 1 for reliabilities.

Target variables. Reading comprehension was evaluated with the Passage Comprehension subtest of the Woodcock Reading Mastery Tests-Revised (Woodcock, 1998), a standardized test

Table 1
Descriptive Statistics for All Variables Organized by Grade (N = 100)

	Grade 3				Grade 4			
Measure	*M*	*SD*	Skew	Reliability	*M*	*SD*	Skew	Reliability
Reading comprehension								
Raw score	33.85	6.82	−0.88	.89[a]	37.33	7.05	−1.06	.92[a]
Standard score	104.43	10.79		.92[b]	102.00	10.99		
Syntactic awareness								
Raw score	7.97	2.71	−0.38	.68[a]	9.02	2.94	−0.75	.74[a]
Morphological awareness								
Raw score	7.91	2.50	−0.24	.73[a]	9.35	2.08	−0.47	.69[a]
Word identification								
Raw score	61.86	11.47	−0.73	.95[a]	69.60	12.34	−0.30	.96[a]
Standard score	102.36	11.93		.97[b]	102.68	14.58		
Word attack								
Raw score	24.28	9.00	−0.42	.93[a]	29.36	8.92	−1.01	.94[a]
Standard score	104.71	12.47		.91[b]	108.99	13.35		
Word reading skills composite								
Z-score	.00	1.00	−0.61	.97[c]	.00	1.00	−0.68	.97[c]
Vocabulary								
Raw score	33.82	5.07	−0.32	.84[d]	36.03	4.88	−0.43	.77[c]
Phonological awareness								
Raw score	8.35	3.31	−0.23	.83[a]	9.84	3.14	−0.60	.88[a]
Nonverbal reasoning								
Raw score	17.65	5.80	−0.41	.92[d]	—	—	—	—
Standard score	101.02	14.57		.94[b]	—	—	—	—
Age								
Months	107.11	3.22	−0.08	—	119.11	3.17	−0.07	—

[a] Sample-specific Cronbach's alpha reliability coefficient. [b] A manual-reported split-half reliability coefficient. [c] Reliability of a composite based on the reliabilities of the two individual components, variances of the two components, and the correlation between them. [d] Sample-specific split-half reliability (with Spearman-Brown corrections).

of reading comprehension that was administered according to the manual's instructions. Participants read short passages, for which they were asked to provide a missing word. An example item is "When the bottlenose dolphin is racing, it arches its back as it leaves the water. It may curve to a height at least twice its own length before striking the ______ again.", with acceptable answers of water, sea, surface, surf. Validity evidence for this and the other published standardized measures is available from the publishers' reports.

Syntactic awareness was measured with a sentence correction task containing three practice items and 14 test items. This task builds on prior sentence correction tasks widely used in the literature (e.g., Bowey, 1986; Willows & Ryan, 1986; Siegel & Ryan, 1988). Children were asked to correct a syntactically incorrect sentence that they heard (e.g., "The teacher the story read to the children"; correct answer: "The teacher read the story to the children."). The majority of these errors were of word order, with the first two including word replacement. All items tap children's awareness of sentence structures broadly, and none of these errors were of morphology. To ensure validity, all sentences included syntactic structures that are produced accurately in the speech of children by 5 years of age (e.g., relative clauses, violation of SVO order). Such productions have been documented in the naturalistic speech of children and confirmed in experimental paradigms (e.g., Brown, 1973; Childers & Tomasello, 2002; Chomsky, 1969; Kidd & Bavin, 2002; Pierce, 1992). All sentences contained one or two clauses, well within the productive ability of children of this age range (Gillam et al., 2010).

Further validity evidence was provided by confirmatory factor analyses with item-level data from this measure and from the vocabulary and morphological awareness measures described below. Results from fitting a three-factor, two-parameter Item Response Theory model (in which the items for each of the three measures loaded on distinct, correlated latent factors) confirmed that the construct tapped by this measure was related to, but distinct from, these other oral language skills in both grades. Specifically, in both grades, the latent correlations between the syntactic awareness and vocabulary factors were moderately sized (Grade 3: .33; Grade 4: .51). These correlations were significantly lower than 1.0, as indicated by Wald parameter constraint tests (Grade 3: Wald χ^2 = 293.86, $df = 1$, $p < .001$; Grade 4: Wald χ^2 = 54.05, $df = 1$, $p < .001$), providing evidence that these two measures tapped distinct factors. Similarly, in both grades, the latent correlations between the syntactic awareness and morphological awareness factors were moderately sized (Grade 3: .42; Grade 4: .46) and significantly lower than 1.0 (Grade 3: Wald χ^2 = 87.66, $df = 1$, $p < .001$; Grade 4: Wald χ^2 = 235.47, $df = 1$, $p < .001$). The latent correlations between morphological awareness and vocabulary were the largest among these correlations, but were also significantly lower than 1.0 in both grades (Grade 3: $r = .84$, Wald χ^2 = 23.65, $df = 1$, $p < .001$; Grade 4: $r = .62$, Wald χ^2 = 63.67, $df = 1$, $p < .001$).

Word reading skills were assessed with Word Identification and Word Attack subtests (Woodcock, 1998) following the manual's instructions. The first evaluates real word reading, and the second nonword reading. Scores for the two tasks were highly correlated in each grade, so our primary analyses were conducted using a composite score (created by averaging within-grade z-scores for the two measures). We conducted additional robustness checks using the individual scores from each measure (see below).

Control measures. Nonverbal rcasoning was assessed with the Matrix Reasoning subtest of the Wechsler Abbreviated Scale of Intelligence (Wechsler, 1999), administered according to manual instructions.

Phonological awareness was assessed with a 13-item phoneme elision task (based on the classic phonological awareness task developed by Rosner & Simon, 1971). Children are asked to say a specified word, and then to say it again but without saying a particular segment or phoneme (e.g., "say *time*. Now say *time* without the /m/").

Vocabulary was measured with the Peabody Picture Vocabulary Test (Dunn & Dunn, 1997). Children indicated which of four pictures displayed on a compact easel best represented a word that was presented orally. We administered every fourth item. This modification reduces testing time and remains highly correlated with full scores (e.g., Deacon, Benere, & Castles, 2012).

Morphological awareness was assessed with a 14-item word analogy task including inflectional and derivational items both with and without a phonological change based on Nunes, Bryant, and Bindman (1997), and as in Deacon, Benere, and Pasquerella (2013) and Deacon et al., (2014). Each item followed the A:B::C:D form (e.g., *run*: *ran*:: *walk*: ___; correct answer *walked*). In addition to evidence from prior studies, validity evidence for this measure was provided by the confirmatory factor analyses described above.

Results

Table 1 presents descriptive statistics for the sample. Raw scores were used in all subsequent analyses. On average, students made significant gains in each skill between Grade 3 and 4 (all $ps < .001$ for paired-samples t tests of mean differences); the magnitudes of these mean gains were substantial and roughly similar for reading comprehension (d = 0.50) and syntactic awareness (d = 0.37). Correlations are shown in Table 2.

Our analyses emphasize effect sizes and confidence intervals, following recent calls to focus on the magnitude and precision of estimates rather than statistical significance alone (e.g., Cumming, 2012). We also compare the magnitude of predictive relations of syntactic awareness to both word reading and reading comprehension with those of other known contributors, such as phonological awareness and vocabulary, respectively. Our analyses here use conservative approaches in controlling for multiple alternative variables. Full information maximum likelihood was used to produce these estimates and all subsequent model estimates.[1] A power analysis supported the use of path analyses with this sample size.[2] Path analyses were conducted using Mplus 7.

[1] This accounts for the minimal missing data on individual measures.

[2] Monte Carlo simulations (Muthén & Muthén, 2002) using the model specifications reported below demonstrated that we could expect acceptably minimal bias in parameter estimates (<5%) and that we had adequate statistical power (above .80) to detect medium sized effects (i.e., standardized regression weights of .3), when the controls also had medium sized effects.

Table 2
Correlations Between Variables in Raw Scores (N = 100)

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. G3 reading comprehension																		
2. G4 reading comprehension	.78																	
3. G3 syntactic awareness	.55	.62																
4. G4 syntactic awareness	.55	.60	.64															
5. G3 morphological awareness	.62	.61	.52	.51														
6. G4 morphological awareness	.68	.64	.54	.63	.72													
7. G3 word identification	.78	.70	.48	.50	.63	.61												
8. G4 word identification	.73	.65	.40	.47	.56	.57	.90											
9. G3 word attack	.65	.58	.36	.44	.51	.59	.88	.88										
10. G4 word attack	.71	.60	.42	.43	.52	.55	.86	.89	.86									
11. G3 word reading Skills composite[a]	.74	.66	.44	.48	.59	.62	.97	.92	.97	.88								
12. G4 word reading Skills composite[a]	.74	.64	.42	.46	.56	.58	.90	.97	.89	.97	.93							
13. G3 vocabulary	.45	.48	.53	.52	.48	.47	.38	.33	.32	.30	.36	.32						
14. G4 vocabulary	.47	.51	.42	.41	.37	.37	.41	.37	.29	.27	.36	.33	.69					
15. G3 phonological awareness	.54	.47	.42	.46	.53	.56	.64	.68	.72	.66	.70	.69	.44	.32				
16. G4 phonological awareness	.53	.55	.47	.46	.53	.60	.61	.61	.61	.57	.62	.61	.43	.40	.66			
17. G3 nonverbal reasoning	.46	.53	.47	.41	.42	.47	.45	.41	.43	.40	.46	.41	.50	.49	.44	.51		
18. G3 age (months)	.05	.08	−.12	−.02	.14	.08	.06	.09	.07	.02	.07	.05	.05	.01	.08	.10	.13	
19. G4 age (months)	.05	.09	−.13	−.04	.14	.06	.06	.07	.06	.01	.06	:04	.08	.02	.08	.09	.13	.99

Note. G = Grade.
[a] Within-grade z-scores used for these composite.

Models Evaluating Mediation by Word Reading Skills in the Relation Between Syntactic Awareness and Reading Comprehension

We first modeled the relations between syntactic awareness and reading comprehension, in particular the extent to which the relation might be mediated by word reading skills in each of Grade 3 and 4. We used multivariate path analysis to fit a partial mediation model (Figure 1b). We compared it to a model in which syntactic awareness and word reading skills both predict reading comprehension, without mediation by word reading skills between syntactic awareness and reading comprehension (Figure 1a). The full specification of the latter model is presented in Figure A1 in the appendix. As shown in Figure A1, each model included control paths from vocabulary, morphological awareness, phonological awareness, nonverbal reasoning, and age to both word reading skills and reading comprehension, to account for the contributions of these established predictors of reading outcomes and their correlations with syntactic awareness. Each model was fitted separately to data from each grade. We compared competing nested models with the Satorra-Bentler χ^2 difference test with robust maximum likelihood estimation, an approach robust to the potential deviations from multivariate normality (e.g., Chou, Bentler, & Satorra, 1991). We emphasize the results from these model comparisons, given our substantive interest in determining which of these a priori theoretical models has better empirical support. Nonetheless, we also evaluated the absolute goodness of fit for the final models across a range of indices, given the relative strengths and weaknesses of different indices (Bollen, 1989), before interpreting the estimates of interest. For individual relations, we employed nonparametric bootstrapped 95% confidence intervals (e.g., Efron & Tibshirani, 1993). This approach is robust to deviations from multivariate normality (given some evidence of skew; see Table 1) and key in estimating the precision of indirect relations (e.g., MacKinnon, Fairchild, & Fritz, 2007).

Model comparisons. Comparisons of the two models displayed in Figure 1 indicated that the model without mediation (1a) was superior to the partial mediation model (1b) in both grades. The additional path in Model 1a did not significantly improve goodness of fit over the more parsimonious Model 1b, in either grade (Grade 3: Satorra-Bentler $\Delta\chi^2 = 0.45$, $\Delta df = 1$, $p = .502$; Grade 4: Satorra-Bentler $\Delta\chi^2 = 0.77$, $\Delta df = 1$, $p = .380$). Confidence intervals indicated that the relations between syntactic awareness and word reading in Model 1a were nonsignificant in both grades (Grade 3: standardized regression coefficients = .07, [−.15, .26]; Grade 4: .10, [−.10, .32]). Similarly, the associated indirect (mediated) relations for syntactic awareness and reading comprehension via word reading were also nonsignificant (Grade 3: .04, [−.07, .14]; Grade 4: .03, [−.05, .11]). In accordance with accepted practice, we removed nonsignificant relations between controls and the reading outcomes to estimate absolute fit statistics (Bollen, 1989); the resulting reduced versions of Model 1b had excellent fit to the data across a range of indices in both grades (Grade 3: RMSEA = 0.00, [0.00, 0.10]; CFI = 1.00; TLI = 1.02; SRMR = 0.037; $\chi^2 = 6.29$, $df = 8$, $p = .615$; Grade 4: RMSEA = 0.00, [0.00, 0.10]; CFI = 1.00; TLI = 1.03; SRMR = 0.04; $\chi^2 = 5.11$, $df = 7$, $p = .646$).

We confirmed with an additional model comparison demonstrating that the direct path between syntactic awareness and reading comprehension was necessary. Including this path significantly improved goodness of fit over a model without this path in both grades (Grade 3: Satorra-Bentler $\Delta\chi^2 = 5.79$, $\Delta df = 1$, $p = .016$; Grade 4: Satorra-Bentler $\Delta\chi^2 = 4.808$, $\Delta df = 1$, $p = .028$).

Magnitude of the relations. Next, we turned to interpreting the magnitude and precision of the individual relations of interest. Syntactic awareness had moderate direct unique relations with reading comprehension in both grades, as shown in Table 3. The unique relation of syntactic awareness to reading comprehension in both Grades 3 and 4 are similar in magnitude to those of two known predictors of reading comprehension: vocabulary knowl-

Table 3

Selected Results for Fitted Multivariate Mediation Models for Concurrent Relations Among Syntactic Awareness, Word Reading Skills, and Reading Comprehension, Controlling for Other Componential Skills ($N = 100$)

		Grade 3		Grade 4	
Path	γ	Standardized coefficient	Bootstrapped 95% CI	Standardized coefficient	Bootstrapped 95% CI
		Direct prediction of reading comprehension			
Syntactic awareness	γ_{21}	.21	[.02, .38]	.21	[.04, .37]
Word reading skills	γ_{22}	.54	[.36, .75]	.33	[.09, .54]
Morphological awareness	γ_{23}	.17	[−.02, .32]	.19	[.02, .37]
Vocabulary	γ_{24}	.06	[−.08, .21]	.18	[.02, .33]
Phonological awareness	γ_{25}	−.07	[−.24, .11]	−.01	[−.22, .19]
Nonverbal reasoning	γ_{26}	.06	[−.10, .21]	.12	[−.05, .28]
Age	γ_{27}	.01	[−.13, .15]	.06	[−.08, .20]
		Indirect prediction of reading comprehension			
Syntactic awareness via Word reading	$\gamma_{11} \times \gamma_{22}$	.04	[−.07, .14]	.03	[−.05, .11]
		Direct prediction of word reading skills			
Syntactic awareness	γ_{11}	.07	[−.15, .26]	.10	[−.12, .32]
Morphological awareness	γ_{13}	.28	[.10, .47]	.26	[.01, .48]
Vocabulary	γ_{14}	−.09	[−.25, .10]	.01	[−.17, .19]
Phonological awareness	γ_{15}	.52	[.35, .66]	.38	[.13, .58]
Nonverbal reasoning	γ_{16}	.11	[−.05, .28]	.05	[−.15, .24]
Age	γ_{17}	−.02	[−.15, .11]	−.01	[−.16, .15]

Note. Model also included covariances among predictors, not shown here. CI = confidence interval.

edge and morphological awareness. Post hoc Wald tests indicated that the standardized coefficients for syntactic awareness in both grades were not significantly different from those for vocabulary knowledge (Grade 3: Wald $\chi^2 = 1.41$, $df = 1$, $p = .235$; Grade 4: Wald $\chi^2 = 0.03$, $df = 1$, $p = .859$) or morphological awareness (Grade 3: Wald $\chi^2 = 0.10$, $df = 1$, $p = .757$; Grade 4: Wald $\chi^2 = 0.01$, $df = 1$, $p = .931$). By contrast, the estimated relations of syntactic awareness to word reading skills are nonsignificant and far smaller than those of phonological awareness, a known predictor of word reading (see Table 3). Post hoc Wald tests indicated that this difference in standardized coefficients was significant in Grade 3 (Wald $\chi^2 = 15.89$, $df = 1$, $p < .001$) and approached significance in Grade 4 (Wald $\chi^2 = 3.47$, $df = 1$, $p = .063$). This suggests that the estimated relations of syntactic awareness to word reading skills are unlikely to be of practical significance.

Together, these models support the conclusion that the contribution of syntactic awareness to reading comprehension is as practically important as the contributions of more established predictors such as vocabulary and morphological awareness. In sharp contrast, the contribution of syntactic awareness to word reading skills, and thereby its contribution to reading comprehension via word reading skills, is nonsignificant. Although these concurrent analyses of observational data cannot support strong causal inferences about mediated effects, they provide more direct tests of hypothesized meditational process than most prior analyses.

Models Evaluating Longitudinal Relations Between Syntactic Awareness and Reading Comprehension

Given the clear direct relations between syntactic awareness and reading comprehension, we then evaluated whether syntactic awareness predicts gains in skill in reading comprehension over time. We used multivariate path analyses to fit the hypothesized autoregressive path analysis model with cross-paths representing bidirectional relations (see Figure 2). The full specification of this model is presented in Figure A2 in the appendix. As shown in Figure A2, this model included control paths from the same reading-related skills included in the models above, measured in Grade 3, and Grade 4 reading comprehension. In addition, these analyses use auto-regressor controls that offer a conservative and powerful estimate of the temporal order of relations between variables (Kenny, 1975); they evaluate whether a given skill accounts for gains in performance on another skill.

Model comparisons. Results indicated that including the cross-path *a* between Grade 3 syntactic awareness and Grade 4 reading comprehension yielded significantly better goodness of fit in comparison to both a model without no cross-paths (Satorra-Bentler $\Delta\chi^2 = 6.93$; $\Delta df = 1$; $p = .010$) and a model with cross-path *b*, but not cross-path *a* (Satorra-Bentler $\Delta\chi^2 = 6.611$; $\Delta df = 1$; $p = .010$). By contrast, including cross-path *b* between Grade 3 reading comprehension and Grade 4 syntactic awareness did not yield significantly better fit relative to a model with path *a* but without path *b* (Satorra-Bentler $\Delta\chi^2 = 1.89$; $\Delta df = 1$; $p = .170$). In accordance with standard practice (Bollen, 1989), we removed nonsignificant control paths from the model without path *b* to estimate absolute fit; fit of the resulting model was excellent across multiple indices (RMSEA = 0.00, [0.00, 0.08]; CFI = 1.00; TLI = 1.05; SRMR = 0.02; $\chi^2 = 2.76$, $df = 6$; $p = .8379$).

Magnitude of the relations. Next, we turned to interpreting the magnitude and precision of the longitudinal relations of interest—the relations between syntactic awareness and gains in reading comprehension and vice versa. Grade 3 syntactic awareness had a moderate, significant unique relation with Grade 4 reading comprehension, after controlling for the auto-regressor (Grade 3 reading comprehension) and Grade 3 word reading skills, vocabulary, phonological awareness, morphological awareness, nonverbal ability, and age, as shown in Table 4. The estimated contribution of Grade 3 syntactic awareness to gains in reading comprehension was very similar in magnitude and precision to those of Grade 3 word reading skills (see Table 4); a post hoc Wald test confirmed that the standardized coefficient for syntactic awareness was not significantly lower than the coefficient for word reading (Wald $\chi^2 = 0.01$, $df = 1$, $p = .935$). This is a key comparison given that word reading skills are an established predictor of reading comprehension (National Institute of Child Health and Human Development, 2000). By contrast, there was no significant unique relation between Grade 3 reading comprehension and Grade 4 syntactic awareness, after accounting for controls.

Models Evaluating Longitudinal Relations Between Syntactic Awareness and Word Reading Skills

To ensure that we fully explored the possibility that there were longitudinal relations between syntactic awareness and word reading skills, we also evaluated whether syntactic awareness predicts gains in word reading skill over time. Results from autoregressive modeling indicated that including a cross-path between Grade 3 syntactic awareness and Grade 4 word reading skills did not significantly improve goodness of fit (Satorra-Bentler $\Delta\chi^2 = 0.003$; $\Delta df = 1$; $p = .956$) and that this path was nonsignificant and trivial in magnitude (standardized coefficient = 0.004, bootstrapped 95% CI = [−0.139, 0.133]).

Table 4
Selected Results for Fitted Autoregressive Multivariate Path Analysis Models for Longitudinal Relations between Grade 3 and Grade 4 Skills (N = 100)

Path		Standardized regression path	Bootstrapped 95% CI
Predicting G4 reading comprehension			
G3 syntactic awareness	γ_{21}	.21	[.04, .38]
G3 reading comprehension	γ_{22}	.43	[.21, .61]
G3 word reading skills	γ_{23}	.20	[.01, .40]
G3 morphological awareness	γ_{24}	.13	[−.04, .31]
G3 vocabulary	γ_{25}	.05	[−.09, .20]
G3 phonological awareness	γ_{26}	−.13	[−.29, .03]
G3 nonverbal reasoning	γ_{27}	.10	[−.06, .24]
G3 age	γ_{28}	.04	[−.10, .18]
Predicting G4 syntactic awareness			
G3 reading comprehension	γ_{12}	.15	[−.04, .38]
G3 syntactic awareness	γ_{11}	.38	[.16, .60]
G3 word reading skills	γ_{13}	.04	[−.20, .30]
G3 morphological awareness	γ_{14}	.05	[−.20, .24]
G3 vocabulary	γ_{15}	.18	[−.02, .35]
G3 phonological awareness	γ_{16}	.09	[−.14, .33]
G3 nonverbal reasoning	γ_{17}	−.02	[−.24, .18]
G3 age	γ_{18}	−.01	[−.19, .18]

Note. Model also included covariances among predictors, not shown here. G = Grade; CI = confidence interval.

Finally, all of the modeling reported on above were confirmed in parallel models replacing the word reading skills composite with word identification or word attack score. In all cases, the same pattern of results emerged when either word attack or word identification was used in place of the word reading composite (details available from authors).

Discussion

Our study was designed to evaluate the relations between syntactic awareness and reading comprehension, assessing the possibility of mediation by word reading, as well as whether syntactic awareness predicts of gains in reading comprehension over time. We did so in a longitudinal study of English-speaking children followed from Grade 3 to 4. First, we show that, at each of Grades 3 and 4, syntactic awareness had a direct, practically important relation to reading comprehension. Furthermore, there was no unique, statistically reliable relations between syntactic awareness and word reading or evidence of mediation. These findings support models of both word reading and reading comprehension development positing direct relations between syntactic awareness and reading comprehension (e.g., Ehri, 2005; Scarborough, 2002), and they conflict with those advocating mediated relations through word reading (e.g., Perfetti et al., 2005). Second, we show that syntactic awareness predicts subsequent gains in reading comprehension between Grades 3 and 4, after controlling for other verbal and nonverbal abilities. These findings point to a potential causal role for syntactic awareness in the development of reading comprehension. Notably, the contribution of syntactic awareness to gains in reading comprehension was similar in size to that of word reading, a strong, established predictor of reading comprehension (NICHD, 2000). Although this evidence is ultimately correlational, it provides valuable support for advancing some causal hypotheses for the role of syntactic awareness in reading development. We first place these results in context of the findings from prior studies and then discuss their theoretical significance and practical importance.

Our findings of a direct relation between syntactic awareness and reading comprehension with no mediation by word reading converge with those of the several studies showing a relation between syntactic awareness and reading comprehension beyond controls for word reading (e.g., Geva & Farnia, 2012). Critically, though, our findings diverge from the two available studies explicitly testing mediation. Tumner and his colleagues found that syntactic awareness predicted reading comprehension indirectly via nonword decoding in two studies with children in Grades 1 and 2 (Tumner et al., 1988; Tumner, 1989); direct relations between syntactic awareness and reading comprehension in these studies were either nonsignificant or not tested. There are several possible explanations for these different results, some of which appear to be more likely than others. In our view, the most likely explanation lies in possible developmental changes; early on there might be indirect relations via word reading at the outset of reading development (as in Tumner et al.'s work) with direct relations emerging later in reading development when word reading is well-established (as in our work). It would be beneficial for future research to investigate this possibility by looking across different stages of reading development. A second, but related, possibility is that the mediation uncovered in studies of young children might be spurious due to the high correlation between word reading and reading comprehension at a young age (e.g., Gough et al., 1996). This is plausible given the lack of direct relations in these studies. Finally, we explored the possibility that differences in the measurement of word reading account for these divergent findings. This is unlikely to be the case, as our further analyses show a parallel pattern of results whether the mediator is nonword decoding (as in Tumner's studies), real word decoding, or a composite of the two. Regardless of the specific explanation for these relations, there is clearly a strong direct relation between syntactic awareness and reading comprehension that does not appear to be mediated by word reading, at least in our study of children in Grades 3 and 4.

Our findings that syntactic awareness predicts gains in reading comprehension converge with those of the one other study in this age range (Farnia & Geva, 2013). As such, we confirm the existence of these relations when measures only target syntactic violations. However, our findings diverge from evidence of no relation to gains in reading comprehension from the one study of much younger children (Blackmore & Pratt, 1997). Taking the results of prior studies and those presented here together, syntactic awareness might predict gains made later, but not early on in the development of reading comprehension. In the reverse direction, and contrary to speculations from Perfetti et al. (2005), we did not find a statistically significant relation between reading comprehension skill and gains in syntactic awareness. This finding is somewhat surprising given the syntactic complexity of the texts encountered by children in this age range (e.g., Williamson et al., 2013). Given the paucity of other data to which we can compare these findings, it would be useful to confirm this pattern in other samples of children, including in other age ranges.

In terms of theory, our findings that syntactic awareness predicts gains in reading comprehension support speculations of the causal role of syntactic awareness put forward by several models of reading comprehension development (e.g., Scarborough, 2002). By finding that syntactic awareness predicts gains in reading comprehension rather than vice versa, we provide evidence for one prerequisite of establishing a causal effect. In addition, by controlling for a robust set of observable third-variables, we account for the most viable alternative hypotheses, though our nonexperimental design does not allow us to account for all possible confounders.

There are two hypothesized mechanisms in current theories for the direct role of syntactic awareness in reading comprehension. Scarborough (2002) described syntactic awareness rather generally as a part of the language comprehension processes that children bring to reading comprehension. Our isolation of the role of syntactic awareness from those of other language skills (specifically vocabulary and morphological awareness) suggests that this is a unique aspect of language skills contributing to children's reading comprehension. Perfetti and his colleagues (2005) articulated a more precise mechanism; syntactic awareness is argued to support children's skill in parsing of sentences into words, which, in turn, facilitates the formation of text level representations that are the foundation of comprehension (see also e.g., Farnia & Geva, 2013; Willows & Ryan, 1986). In our view, both suggested mechanisms are in need of further empirical evaluation, with specific attention to the kinds of syntactic complexity that children encounter in text (e.g., Williamson et al., 2013). It seems plausible in our view that both such mechanisms are at work. Children likely need to be aware of syntactic forms in the oral language to be able to understand these forms in the written language. This awareness could then give them a means to break down complex sentences, parsing them into more manageable chunks which can then be used to build comprehension. As such, both of these mechanisms are worthy of targeted empirical investigation.

Our finding that word reading does not mediate the relation between syntactic awareness and reading comprehension converges with predictions from models of word reading (e.g., Ehri, 2015) and diverges from those of one prominent model of reading comprehension (Perfetti et al., 2005). Two mechanisms have been put forward to explain the relation between syntactic awareness and word reading. Perfetti (e.g., Perfetti & Hart, 2002) suggested one possible mechanism, conceptualizing syntax as being a part of lexical representations. Tumner (1989) put forward a second mechanism, suggesting that syntactic awareness permits children to combine partial decoding with contextual cues to arrive at the correct word pronunciation. Our findings of no statistically detectable unique role for syntactic awareness in word reading counters both views; in our opinion, these findings need to be taken seriously as they were remarkably consistent across analyses at two different grade levels and longitudinally (see also e.g., Bowey, 1986). Nevertheless, we think that, as with all scientific investigations, these findings need to be considered in light of the specific context in which they emerge. Our analyses focused on the unique role for syntactic awareness. It is possible that syntax is so embedded within high quality lexical representations that there is no unique role that can separated from their phonological and semantic dimensions; the latter two may have been captured by controls for phonological awareness and vocabulary in our analyses. This possibility is supported by the intercorrelations between measures. Further, we evaluated the role of syntactic awareness in word reading, as captured by standardized measures of single words and nonwords. We did not assess children's use of context, which would be the most direct test of Tumner's speculations. Browne Rego and Bryant (1993) demonstrated such a relation in a study with a small group of children in Grade 1; it would be important to evaluate these relations with older children (see Mimeau, Laroche, & Deacon, 2016 for a recent attempt). Such explicit theoretical tests would further clarify our findings of no mediated relation between syntactic awareness and reading comprehension via word reading.

In terms of educational implications, our findings suggest that syntactic awareness is one skill that appears to support children's comprehension of texts. This would point to the possibility that teaching children about syntactic structures could lead to improvements in their reading comprehension. In our review of reading comprehension interventions to date, we noticed that these have tended to focus on teaching vocabulary and specific comprehension strategies (e.g., RAND, 2002). Our findings suggest that it might be useful to add instruction on syntax or sentence-level information, as has been included in some recent interventions (e.g., Clarke, Snowling, Truelove, & Hulme, 2010). Specifically, the relation between syntactic awareness and gains in reading comprehension over time suggest that such targeted instruction could have strong knock-on effects on children's ability to understand texts, particularly given their increasing complexity (e.g., Snow, 2010). Clearly, such a speculation needs further empirical testing. Our findings also point to what might not be effective in classroom instruction. Tumner (1989) and others speculated that syntactic awareness might allow children to combine partial decoding with contextual cues to arrive at the correct word pronunciation (e.g., Muter & Snowling, 1998; Tumner et al., 1988). In our own study, the absence of a unique relation between syntactic awareness and word reading skill suggests that this is not the case. These findings resonate with evidence that it is poor, rather than good readers who rely on sentence context to support their reading of individual words (for reviews, see Ehri, 2015; Share, 2008). Together, these studies suggest that children should not be taught to rely on sentence-level information to guess at individual words that they do not know, a practice that endures in classrooms (e.g., Calkins, Ehrenworth, & Lehman, 2012; Fountas & Pinnell, 2010). Awareness of syntactic information could provide a useful way to help children deal with complex texts, specifically by supporting children's understanding of texts, rather their reading of individual words.

To further situate this work, it is useful to consider the rising linguistic demands of texts in the classroom and workplace. Texts in the workplace are far more complex than they have been in the past (Adams, 2009). Specifically, it is the syntactic complexity in texts that is challenging. Syntactic complexity increases more sharply across grade levels than any other language feature (Graesser, McNamara, & Kulikowich, 2011), with complex sentences being the single-most important feature that makes texts harder to read (e.g., Klare, 1976; Stenner & Swartz, 2012). There is growing interest in developing evidence-based knowledge of how to prepare children to meet these reading demands (e.g., Common Core State Standards Initiative, 2010; Snow, 2010). Our findings point to the promise of future research into how instruction can increase children's syntactic awareness and thereby improve their reading comprehension outcomes.

As with all studies, we need to bear in mind specific limitations to our study. One lies in our choice of a reading comprehension measure. Passage comprehension is a well-established measure of reading comprehension that is ideally suited to examine individual differences over time. Notable, in contrast to some other measures of reading comprehension, performance on passage comprehension is more highly correlated with word reading skills (Keenan, Betjemann, & Olson, 2008). Although this likely reduced our ability to detect mediation by word reading skills, the consistency of results across our re-analyses with alternative measure of word reading go some distance in allaying these concerns. Another limitation lies in control measures that were included. We controlled for several language skills, specifically vocabulary and morphological awareness. Another relevant control might be listening comprehension because syntax operates at the sentence level. That said, we are confident that our measurement of vocabulary provided a critical control. In the one available study looking at children in this age range who were learning to read a similar orthography (Dutch), vocabulary, and not listening comprehension, predicted unique variance in gains in reading comprehension (Verhoeven & Van Leeuwe, 2008). However, there are no available comparable studies with English-speaking children. Clearly new research needs to fill the empirical gap in contrasting vocabulary, listening comprehension and both syntactic and morphological awareness simultaneously as predictors of gains in reading comprehension. Such research could also include memory as a control, given the length of the sentences the children are required to remember during the syntactic awareness task (e.g., Cain, 2007). Finally, the examination of relations with two waves of data over a single 12-month period prevents us from testing longitudinal patterns of mediation; future longitudinal models incorporating three or more waves of data with parallel measures may provide stronger tests of indirect effects via word reading. The design of the current study also precludes our ability to evaluate a possible cumulative cycle; syntactic awareness could predict gains in reading comprehension, which could, in turn predict subsequent gains in syntactic awareness (see also Scarborough, 2002). The inclusion of an auto-regressor means that we might have controlled for this prior developmental relation. Nonetheless, longitudinal studies over a longer time period could investigate this possibility more directly.

Beyond these methodological limitations, there is a conceptual challenge faced by all studies of metalinguistic skills: demonstrating that performance reflects awareness, rather than knowledge of linguistic features. To deal with this challenge, Bowey (1986) ensured that all errors in the task were documented in the speech of much younger children increasing confidence that the children in her study should already have the required knowledge. We chose to build on this approach; however, instead of relying on errors, we turned to evidence of accuracy of production. We chose syntactic structures for which we have good evidence that much younger children can produce, specifically by 5 years of age, a full three years younger than the children tested in our study. We did so to give us confidence that differences in levels of performance on our syntactic awareness task reflect different levels of awareness, rather than productive knowledge of syntax. Nevertheless, we think that there is a good deal more work to be done on this front, particularly building on earlier similar work in phonology (see, e.g., Ramus, & Szenkovits, 2008 for a review). For example, manipulations of the degree of awareness required in a task would be useful. We point to this as an important direction for future research.

In summary, we have shown that children's syntactic awareness uniquely predicts their concurrent levels and future gains in reading comprehension. Contrary to the predictions of some models of reading comprehension, there is no evidence that syntactic awareness predicts children's word reading. These findings encourage us to retain an emphasis on syntactic awareness in theories as a direct determiner specifically of skill in reading comprehension.

References

Adams, M. J. (2009). The challenge of advanced texts: The interdependence of reading and learning. In E. H. Hiebert (Ed.), *Reading more, reading better* (pp. 1–38). New York, NY: Guilford Press Publications.

Blackmore, A. M., & Pratt, C. (1997). Grammatical awareness and reading in Grade 1 children. *Merrill-Palmer Quarterly, 43,* 567–590.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York, NY: Wiley. http://dx.doi.org/10.1002/9781118619179

Bowey, J. A. (1986). Syntactic awareness in relation to reading skill and ongoing reading comprehension monitoring. *Journal of Experimental Child Psychology, 41,* 282–299. http://dx.doi.org/10.1016/0022-0965(86)90041-X

Bowey, J. A. (1994). Grammatical awareness and learning to read: A critique. In E. Assink (Ed.), *Literacy acquisition and social context* (pp. 122–149). London, UK: Harvester Wheatsheaf/Prentice Hall.

Bradley, L., & Bryant, P. E. (1983). Categorizing sounds and learning to read—a causal connection. *Nature, 301,* 419–421. http://dx.doi.org/10.1038/301419a0

Brimo, D., Apel, K., & Fountain, T. (2017). Examining the contributions of syntactic awareness and syntactic knowledge to reading comprehension. *Journal of Research in Reading, 40,* 57–74. http://dx.doi.org/10.1111/1467-9817.12050

Brown, R. (1973). *A first language: The early stages.* Cambridge,MA: Harvard University Press. http://dx.doi.org/10.4159/harvard.9780674732469

Browne Rego, L., & Bryant, P. (1993). The connection between phonological syntactic and semantic skills and children's reading and spelling. *European Journal of Psychology of Education, 8,* 235–246. http://dx.doi.org/10.1007/BF03174079

Cain, K. (2007). Syntactic awareness and reading ability: Is there any evidence for a special relationship? *Applied Psycholinguistics, 28,* 679–694. http://dx.doi.org/10.1017/S0142716407070361

Calkins, L., Ehrenworth, M., & Lehman, C. (2012). *Pathways to the common core: Accelerating achievement.* Portsmouth, NH: Heinemann.

Chafe, W. (1985). Linguistic differences produced by differences between speaking and writing. In D. Olson, N. Torrance, & A. Hildyard (Eds.), *Literacy, language, and learning: The nature and consequences of reading and writing* (pp. 105–122). Cambridge, UK: Cambridge University Press.

Chall, J. S. (1983). *Stages of reading development.* New York, NY: McGraw-Hill.

Childers, J. B., & Tomasello, M. (2002). Two-year-olds learn novel nouns, verbs, and conventional actions from massed or distributed exposures. *Developmental Psychology, 38,* 967–978. http://dx.doi.org/10.1037/0012-1649.38.6.967

Chomsky, C. (1969). *The acquisition of syntax in children from 5 to 10.* Cambridge, MA: The MIT Press.

Chou, C. P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 44,* 347–357. http://dx.doi.org/10.1111/j.2044-8317.1991.tb00966.x

Clarke, P. J., Snowling, M. J., Truelove, E., & Hulme, C. (2010). Ameliorating children's reading-comprehension difficulties: A randomized controlled trial. *Psychological Science, 21,* 1106–1116. http://dx.doi.org/10.1177/0956797610375449

Common Core State Standards Initiative. (2010). *Supplemental information for Appendix A of the Common Core standards for English language arts and literacy: New research on text complexity.* Retrieved from http://www.corestandards.org/other-resources/

Crystal, D. (1987). *The Cambridge encyclopedia of language.* New York, NY: Cambridge University Press.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York, NY: Routledge.

Deacon, S. H., Benere, J., & Castles, A. (2012). Chicken or egg? Untangling the relationship between orthographic processing skill and reading accuracy. *Cognition, 122,* 110–117. http://dx.doi.org/10.1016/j.cognition.2011.09.003

Deacon, S. H., Benere, J., & Pasquarella, A. (2013). Reciprocal relationship: Children's morphological awareness and their reading accuracy across grades 2 to 3. *Developmental Psychology, 49,* 1113–1126. http://dx.doi.org/10.1037/a0029474

Deacon, S. H., Kieffer, M., & Laroche, A. (2014). The relation between morphological awareness and reading comprehension: Evidence from mediation and longitudinal models. *Scientific Studies of Reading, 18,* 432–451. http://dx.doi.org/10.1080/10888438.2014.926907

Deacon, S. H., & Kirby, J. R. (2004). Morphological awareness: Just "more phonological"? The roles of morphological and phonological awareness in reading development. *Applied Psycholinguistics, 25,* 223–238. http://dx.doi.org/10.1017/S0142716404001110

Demont, E., & Gombert, J. E. (1996). Phonological awareness as a predictor of recoding skills and syntactic awareness as a predictor of comprehension skills. *British Journal of Educational Psychology, 66,* 315–332. http://dx.doi.org/10.1111/j.2044-8279.1996.tb01200.x

Dunn, L. M., & Dunn, L. C. (1997). *PPVT-III: Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Education Quality and Accountability Office. (2015). *Primary division language reading book.* Retrieved from http://www.eqao.com/en/assessments/primary-division/Pages/example-assessment-materials-2015.aspx

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York, NY: Chapman & Hall. http://dx.doi.org/10.1007/978-1-4899-4541-9

Ehri, L. C. (1976). Word learning in beginning readers and pre-readers: Effects of form class and defining contexts. *Journal of Educational Psychology, 68,* 832–842. http://dx.doi.org/10.1037/0022-0663.68.6.832

Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading, 9,* 167–188. http://dx.doi.org/10.1207/s1532799xssr0902_4

Ehri, L. C. (2015). How children learn to read words. In A. Pollatsek & R. Treiman (Eds.), *The Oxford handbook of reading* (pp. 293–310). New York, NY: Oxford University Press.

Farnia, F., & Geva, E. (2013). Growth and predictors of change in English language learners' reading comprehension. *Journal of Research in Reading, 36,* 389–421.

Foorman, B. R., Petscher, Y., & Bishop, M. (2012). The incremental variance of morphological knowledge to reading comprehension in Grades 3–10 beyond prior reading comprehension, spelling, and text reading efficiency. *Learning and Individual Differences, 22,* 792–798. http://dx.doi.org/10.1016/j.lindif.2012.07.009

Fountas, I., & Pinnell, G. S. (2010). *The continuum of literacy learning, Grades PreK-2: A guide to teaching* (2nd ed.). Portsmouth, NH: Heinemann.

Geva, E., & Farnia, F. (2012). Developmental changes in the nature of language proficiency and reading fluency paint a more complex view of reading comprehension in ELL and EL1. *Reading and Writing, 25,* 1819–1845. http://dx.doi.org/10.1007/s11145-011-9333-8

Gillam, R., Marquardt, T. P., & Martin, F. N. (2010). *Communication sciences and disorders: From science to clinical practice.* Boston, MA: Jones & Bartlett.

Gottardo, A., Stanovich, K. E., & Siegel, L. S. (1996). The relationship between phonological sensitivity, syntactic processing and verbal working memory in the reading performance of third grade children. *Journal of Experimental Child Psychology, 63,* 563–582. http://dx.doi.org/10.1006/jecp.1996.0062

Gough, P. B., Hoover, W. A., & Peterson, C. L. (1996). Some observations on a simple view of reading. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 1–13). Mahwah, NJ: Erlbaum Publishers.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix providing multilevel analyses of text characteristics. *Educational Researcher, 40,* 223–234. http://dx.doi.org/10.3102/0013189X11413260

Hollingshead, A. B. (1957). *Two factor index of social position.* Unpublished manuscript, Department of Sociology, Yale University, New Haven, CT.

Holsgrove, J. V., & Garton, A. F. (2006). Phonological and syntactic processing and the role of working memory in reading comprehension among secondary school students. *Australian Journal of Psychology, 58,* 111–118. http://dx.doi.org/10.1080/00049530600730476

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12,* 281–300. http://dx.doi.org/10.1080/10888430802132279

Kenny, D. (1975). Cross-lagged panel correlation: A test for spuriousness. *Psychological Bulletin, 82,* 887–903. http://dx.doi.org/10.1037/0033-2909.82.6.887

Kidd, E., & Bavin, E. L. (2002). English-speaking children's comprehension of relative clauses: Evidence for general-cognitive and language-specific constraints on development. *Journal of Psycholinguistic Research, 31,* 599–617. http://dx.doi.org/10.1023/A:1021265021141

Klare, G. R. (1976). A second look at the validity of readability formulas. *Journal of Reading Behavior, 8,* 129–152.

Klauda, S. L., & Guthrie, J. T. (2008). Relationships of three components of reading fluency to reading comprehension. *Journal of Educational Psychology, 100,* 310–321. http://dx.doi.org/10.1037/0022-0663.100.2.310

Kruk, R. S., & Bergman, K. (2013). The reciprocal relations between morphological processes and reading. *Journal of Experimental Child Psychology, 114,* 10–34. http://dx.doi.org/10.1016/j.jecp.2012.09.014

Kuo, L., & Anderson, R. C. (2006). Morphological awareness and learning to read: A cross language perspective. *Educational Psychologist, 41,* 161–180. http://dx.doi.org/10.1207/s15326985ep4103_3

Low, P., & Siegel, L. (2005). A comparison of the cognitive processes underlying reading comprehension in native English and ESL speakers. *Written Language and Literacy, 8,* 207–231.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58,* 593–614. http://dx.doi.org/10.1146/annurev.psych.58.110405.085542

Mimeau, C., Laroche, A., & Deacon, S. H. (2016). *The relation between syntactic awareness and contextual facilitation in word reading: What is the role of semantics?* Manuscript submitted for publication.

Muter, V., Hulme, C., Snowling, M. J., & Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: Evidence from a longitudinal study. *Developmental Psychology, 40,* 665–681. http://dx.doi.org/10.1037/0012-1649.40.5.665

Muter, V., & Snowling, M. (1998). Concurrent and longitudinal predictors of reading: The role of metalinguistic and short term memory skills. *Reading Research Quarterly, 33,* 320–337. http://dx.doi.org/10.1598/RRQ.33.3.4

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9,* 599–620. http://dx.doi.org/10.1207/S15328007SEM0904_8

Nagy, W. (2007). Metalinguistic awareness and the vocabulary-comprehension connection. In R. K. Wagner, A. E. Muse, & K. R. Tannenbaum (Eds.), *Vocabulary acquisition: Implications for reading comprehension* (pp. 52–77). New York, NY: Guilford Press.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00–4769). Washington, DC: U.S. Government Printing Office.

Nunes, T., Bryant, P., & Bindman, M. (1997). Morphological spelling strategies: Developmental stages and processes. *Developmental Psychology, 33,* 637–649. http://dx.doi.org/10.1037/0012-1649.33.4.637

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11,* 357–383. http://dx.doi.org/10.1080/10888430701530730

Perfetti, C. A., Beck, I., Bell, L., & Hughes, C. (1987). Phonemic knowledge and learning to read are reciprocal: A longitudinal study of first grade children. *Merrill-Palmer Quarterly, 33,* 283–319.

Perfetti, C. A., & Hart, L. (2002). The lexical basis of comprehension skill. In D. S. Gorfein & D. S. Gorfein (Eds.), *On the consequences of meaning selection: Perspective on resolving lexical ambiguity* (pp. 67–86). Washington, DC: American Psychological Association.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Malden, MA: Blackwell Publishing. http://dx.doi.org/10.1002/9780470757642.ch13

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18,* 22–37. http://dx.doi.org/10.1080/10888438.2013.827687

Pierce, A. (1992). *Language acquisition and syntactic theory: A comparative analysis of French and English child grammars.* Dordrecht, The Netherlands: Kluwer. http://dx.doi.org/10.1007/978-94-011-2574-1

Ramus, F., & Szenkovits, G. (2008). What phonological deficit? *The Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 61,* 129–141. http://dx.doi.org/10.1080/17470210701508822

RAND Reading Study Group. (2002). *Reading for understanding: Toward a reading and development program in reading comprehension.* Santa Monica, CA: RAND.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124,* 372–422. http://dx.doi.org/10.1037/0033-2909.124.3.372

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2,* 31–74. http://dx.doi.org/10.1111/1529-1006.00004

Rosner, J., & Simon, D. (1971). The auditory analysis test: An initial report. *Journal of Learning Disabilities, 4,* 384–392. http://dx.doi.org/10.1177/002221947100400706

Scarborough, H. S. (2002). Connecting early language and literacy to later (dis)abilities. Evidence, theory and practice. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (pp. 97–110). New York, NY: Guilford Press.

Share, D. L. (2008). Orthographic learning, phonological recoding, and self-teaching. In R. V. Kail (Ed.), *Advances in child development and behavior* (Vol. 36, pp. 31–82). San Diego, CA: Elsevier Academic Press.

Siegel, L. S., & Ryan, E. B. (1988). Development of grammatical sensitivity, phonological and short-term memory skills in normally achieving and learning disabled children. *Developmental Psychology, 24,* 28–37. http://dx.doi.org/10.1037/0012-1649.24.1.28

Snow, C. E. (2010). Academic language and the challenge of reading for learning about science. *Science, 328,* 450–452. http://dx.doi.org/10.1126/science.1182597

Stenner, A. J., & Swartz, C. (2012, April). *A causal Rasch model for understanding comprehension in the context of reader-text-task.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Tumner, W. E. (1989). The role of language-related factors in reading disability. In D. Shankweiler & I. Y. Liberman (Eds.), *Phonology and reading disability: Solving the reading puzzle* (pp. 91–131). Ann Arbor, MI: University of Michigan Press.

Tumner, W., Herriman, M., & Nesdale, A. (1988). Metalinguistic abilities and beginning reading. *Reading Research Quarterly, 23,* 134–158. http://dx.doi.org/10.2307/747799

Verhoeven, L., & Perfetti, C. (2008). Advances in text comprehension: Model, process and development. *Applied Cognitive Psychology, 22,* 293–301. http://dx.doi.org/10.1002/acp.1417

Verhoeven, L., & Van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology, 22,* 407–423. http://dx.doi.org/10.1002/acp.1414

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence.* San Antonio, TX: PsychCorp.

Williamson, G. L., Fitzgerald, J., & Stenner, A. J. (2013). The common core state standards' quantitative text complexity trajectory: Figuring out how much complexity is enough. *Educational Researcher, 42,* 59–69. http://dx.doi.org/10.3102/0013189X12466695

Willows, D. M., & Ryan, E. B. (1986). The development of grammatical sensitivity and its relationship to early reading achievement. *Reading Research Quarterly, 21,* 253–266. http://dx.doi.org/10.2307/747708

Woodcock, R. W. (1998). *Woodcock Reading Mastery Tests—Revised/normative update.* Circle Pines, MN: American Guidance Service.

Appendix

Items in the Syntactic Awareness Task

#	Item presented to child	Correct response(s)
1	The boy found the book what you lost.	The boy found the book that you lost.
2	John gave the crayon for Mary.	John gave the crayon to Mary.
3	Peter goes sometimes to church.	Peter sometimes goes to church. OR Peter goes to church sometimes.
4	The girl lost her money who lives across the road.	The girl who livcs across the road lost her money. OR The girl who lost her money lives across the road.
5	The teacher the story read to the children.	The teacher read the story to the children.
6	She will be angry if you will break it.	She will be angry if you break it.
7	The boy forgot his uniform who plays baseball.	The boy who plays baseball forgot his uniform. OR The boy who forgot his uniform plays baseball.
8	Found in the ocean are whales.	Whales are found in the ocean.
9	Interested in music Mary wasn't.	Mary wasn't interested in music.
10	She swims not.	She doesn't swim.
11	Herself likes to dress Celina.	Celina likes to dress herself.
12	Were eaten by the dog the cookies.	The cookies were eaten by the dog.
13	The boy gave the ball who was crying to the baby.	The boy gave the ball to the baby who was crying.
14	From the library were stolen the books.	The books were stolen from the library.

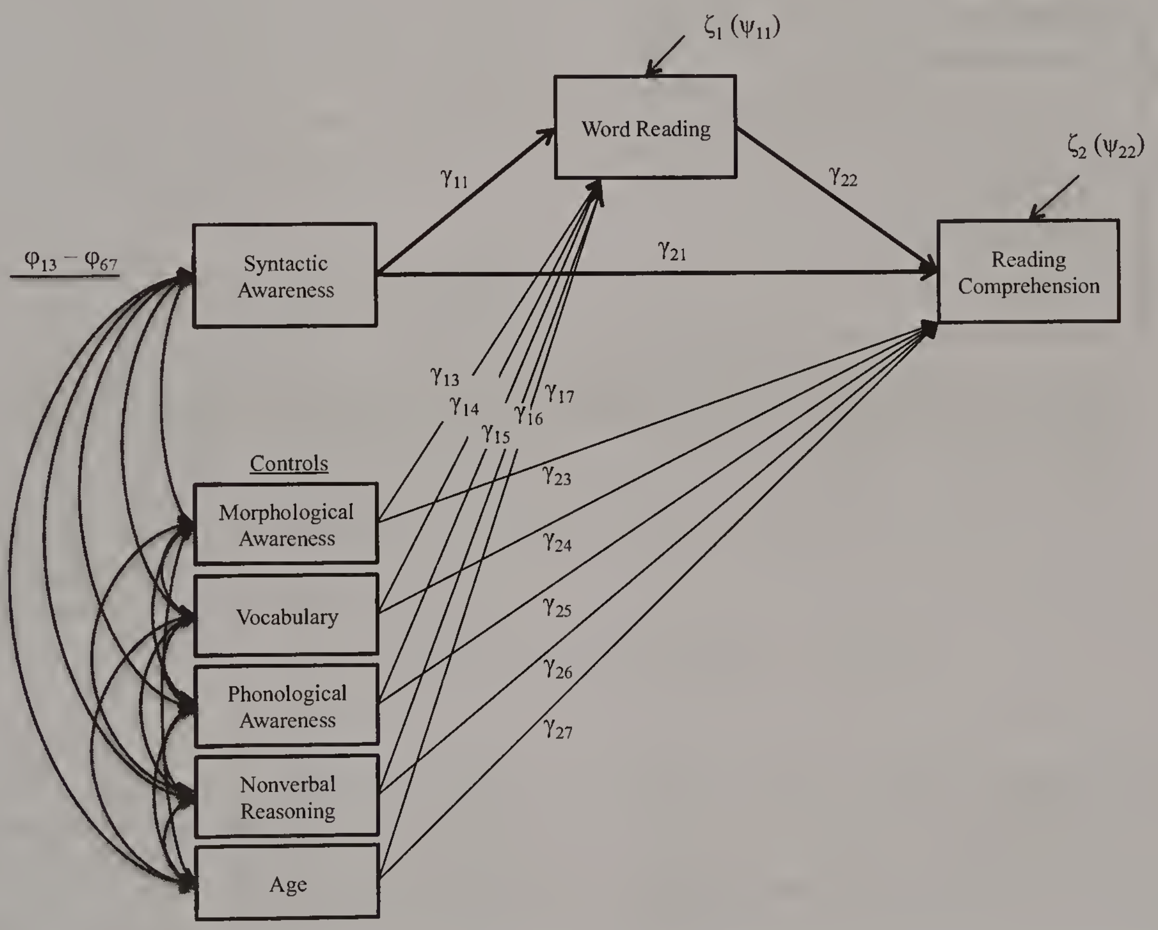


Figure A1. Full specification of the hypothesized path analysis model for the concurrent relations among syntactic awareness and reading comprehension, fully mediated by word reading, and accounting for control paths from established reading-related skills.

(*Appendix continues*)

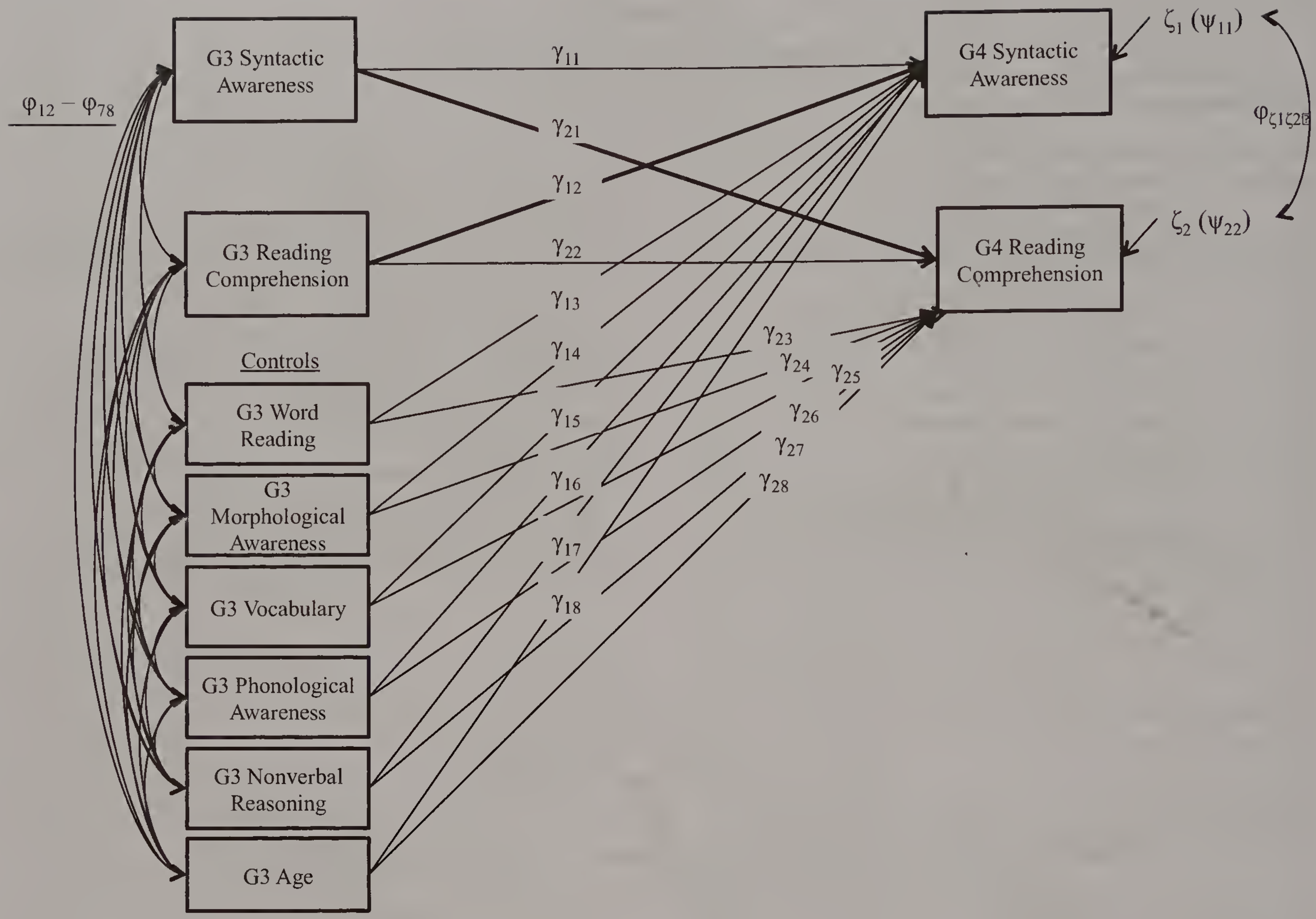


Figure A2. Full specification of hypothesized path analysis model for the longitudinal relations among syntactic awareness and reading comprehension in Grades 3 and 4, accounting for control paths from established reading-related skills in Grade 3.

Received July 26, 2016
Revision received January 20, 2017
Accepted February 6, 2017 ■

Journal of Educational Psychology
2018, Vol. 110, No. 1, 87–101

© 2017 American Psychological Association
0022-0663/18/$12.00 http://dx.doi.org/10.1037/edu0000208

Assessing Formal Knowledge of Math Equivalence Among Algebra and Pre-Algebra Students

Emily R. Fyfe
Indiana University

Percival G. Matthews
University of Wisconsin–Madison

Eric Amsel
Weber State University

Katherine L. McEldoon
Tennessee Department of Education, Nashville, Tennessee

Nicole M. McNeil
University of Notre Dame

A central understanding in mathematics is knowledge of *math equivalence*, the relation indicating that 2 quantities are equal and interchangeable. Decades of research have documented elementary-school (ages 7 to 11) children's (mis)understanding of math equivalence, and recent work has developed a construct map and comprehensive assessments of this understanding. The goal of the current research was to extend this work by assessing whether the construct map of math equivalence knowledge was applicable to middle school students and to document differences in formal math equivalence knowledge between students in pre-algebra and algebra. We also examined whether knowledge of math equivalence was related to students' reasoning about an algebraic expression. In the study, 229 middle school students (ages 12 to 16) completed 2 forms of the math equivalence assessment. The results suggested that the construct map and associated assessments were appropriate for charting middle school students' knowledge and provided additional empirical support for the link between understanding of math equivalence and formal algebraic reasoning.

Educational Impact and Implications Statement

One of the bedrocks of algebraic thinking is formal knowledge of math equivalence, which is the idea that 2 sides of an equation are equal and interchangeable. In the present study, we sought to validate a measure of this knowledge in middle school students. Students in pre-algebra and algebra classes were successful on some items but still struggled with others (such as explicitly defining the equal sign or reasoning about operations on both sides of the equal sign, e.g., explaining why 89 + 44 = 87 + 46 is true). We also found that performance on the measure was related to students' formal understanding of algebraic expressions. Our results highlight the importance of measuring formal knowledge of math equivalence beyond elementary school, particularly given its critical link to algebra.

Keywords: algebra, mathematical equivalence, measurement development, mathematics

One of the bedrocks of early algebraic thinking is knowledge of *math equivalence*, the relation indicating that two quantities are equal and interchangeable (e.g., Baroody & Ginsburg, 1983; Carpenter, Franke, & Levi, 2003; Kieran, 1981; MacGregor & Stacey, 1997). Unfortunately, much research has documented students' (mis)understanding of math equivalence in symbolic form (e.g., Behr, Erlwanger, & Nichols, 1980; Knuth, Alibali, McNeil, Weinberg, & Stephens, 2005; Li, Ding, Capraro, & Capraro, 2008; Lindvall & Ibarra, 1980; Powell & Fuchs, 2010; Weaver, 1973). This research underscores the need for assessments that can both

This article was published Online First May 4, 2017.

Emily R. Fyfe, Department of Psychological and Brain Sciences, Indiana University; Percival G. Matthews, Department of Educational Psychology, University of Wisconsin-Madison; Eric Amsel, Department of Psychology, Weber State University; Katherine L. McEldoon, Tennessee Department of Education, Nashville, Tennessee; Nicole M. McNeil, Department of Psychology, University of Notre Dame.

The authors thank Daniel Bolt for helpful guidance on the methodological aspects of this article.

Support for this research was provided in part by Institute of Education Sciences, U.S. Department of Education Training Grant R305B130007 as part of the Wisconsin Center for Education Research Postdoctoral Training Program.

Correspondence concerning this article should be addressed to Emily R. Fyfe, Department of Psychological and Brain Sciences, Indiana University, 1101 East, 10th Street, Bloomington, IN 47405. E-mail: efyfe@indiana.edu

track students' formal knowledge of math equivalence and serve as valid and reliable outcome measures for intervention work. Recent research has developed comprehensive assessments of this understanding among elementary schoolchildren, ages 7 to 11 (Matthews, Rittle-Johnson, McEldoon, & Taylor, 2012; Rittle-Johnson, Matthews, Taylor, & McEldoon, 2011). The primary goal of the current study was to investigate whether the assessments could reliably and validly measure formal knowledge of math equivalence among an older cohort of students, ages 12 to 16, and to report differences in math equivalence knowledge between students in pre-algebra and algebra classes. We also examined how performance on the equivalence measure was related to formal reasoning about a specific algebraic expression.

Knowledge of Math Equivalence

Math equivalence is a broad construct and a formal understanding of it encompasses a number of related components (e.g., Falkner, Levi, & Carpenter, 1999; Charles, 2005; Kieran, 1981; McNeil & Alibali, 2005). One component is a relational understanding of the equal sign (i.e., knowing that values on either side of the equal sign need to be the same amount). However, other key components include correctly identifying the sides of an equation, noticing relations within equations, encoding equations in their entirety (e.g., noticing the location and order of operators, quantities, and the equal sign), generating a strategy for equalizing the two sides of an equation (e.g., solving for an unknown), and knowing that a quantity can be represented in many equal and interchangeable ways (e.g., knowing that 7 can be represented as 3 + 4, 8 – 1, 2 + 1 + 4, 7 × 1, 14/2, etc.). Of course, children can have an informal understanding of math equivalence without mapping that understanding to the formal symbols (e.g., Mix, 1999; Sherman & Bisanz, 2009). In the current study, we were specifically interested in student's formal understanding of math equivalence—though we use the term "math equivalence" for conciseness.

Math equivalence is considered a "big idea" in mathematics because it is both critical to learning mathematics and because it facilitates an understanding of mathematics as a coherent whole, rather than as a set of unrelated concepts and procedures (Charles, 2005; National Council of Teachers of Mathematics [NCTM], 2000). Further, math equivalence is considered a critical prerequisite to formal algebraic understanding (e.g., Jacobs, Franke, Carpenter, Levi, & Battey, 2007; Kieran, 1981; Knuth, Stephens, McNeil, & Alibali, 2006; MacGregor & Stacey, 1997). Accordingly, the Common Core State Standards recognize the importance of math equivalence and prescribe that children as early as first grade should be able to understand the relational meaning of the equal sign, to determine unknown numbers in equations (e.g., 12 = 5 + __), and to judge equations as true or false (e.g., 6 = 6, 5 + 2 = 2 + 5; National Governors Association Center for Best Practices and the Council of Chief State School Officers, 2010).

Unfortunately, decades of research in psychology and mathematics education indicate that many elementary schoolchildren (ages 7 to 11) in the United States struggle to understand math equivalence in symbolic form (Behr et al., 1980; Falkner et al., 1999; Fyfe, McNeil, & Borjas, 2015; Fyfe, Rittle-Johnson, & DeCaro, 2012; McNeil & Alibali, 2005; Perry, 1991; Powell & Fuchs, 2010; Renwick, 1932; Rittle-Johnson, 2006; Rittle-Johnson & Alibali, 1999; Seo & Ginsburg, 2003; Weaver, 1973). The broad issue is one of a "cognitive gap" between arithmetic and algebra (Herscovics & Linchevski, 1994; van Amerom, 2003). Students' experiences with arithmetic often lead them to view equations operationally as computational processes to be carried out, rather than structurally as arguments whose truths can be evaluated by the products of those processes (e.g., Kieran, 1981; Linchevski & Herscovics, 1996; Sfard & Linchevski, 1994). For example, children often view a problem like 3 + 4 = 5 + __ as a signal to carry out a procedure rather than as two related sides or "objects," (3 + 4) and (5 + __), whose substitutability is expressed by the equal sign. This "gap" leads to specific errors on a variety of problems assessing knowledge of math equivalence. For example, children often provide an operational definition of the equal sign—inferring that it means "get the answer" or "the total" (e.g., Baroody & Ginsburg, 1983; Behr et al., 1980; McNeil & Alibali, 2005). When solving problems with operations on both sides of the equal sign (e.g., 3 + 4 = 5 + __), children often fall into entrenched patterns of calculation and add up all the numbers (12) to write the total in the blank (e.g., Alibali, 1999; Falkner et al., 1999; Li et al., 2008). Children also tend to reject problems that are not in standard operations-equals-answer format, claiming that problems such as 8 = 5 + 3 are backward or false (e.g., Behr et al., 1980; Falkner et al., 1999; Li et al., 2008; Molina & Ambrose, 2008; Rittle-Johnson & Alibali, 1999; Seo & Ginsburg, 2003).

The Construct Map for Math Equivalence

Recent work (e.g., Matthews et al., 2012; Rittle-Johnson et al., 2011) has made strides in elucidating math equivalence knowledge as a construct and creating assessments that pool items and tasks from previous research in both psychology and math education (e.g., Baroody & Ginsburg, 1983; Behr et al., 1980; Carpenter, Franke, & Levi, 2003; Jacobs et al., 2007; Li et al., 2008; McNeil & Alibali, 2005; Perry, 1991; Rittle-Johnson & Alibali, 1999; Sherman & Bisanz, 2009; Steinberg et al., 1991; Weaver, 1973). This type of work is critical in order to document the relative difficulties of all the different types of items used in the literature to measure the same construct and to better understand the order in which children's knowledge is constructed. The construct map put forth by Rittle-Johnson et al. (2011) is shown in Table 1. It was derived after an extensive review of the literature on children's math equivalence knowledge.

The construct map contains four levels of increasing sophistication of knowledge. Although the map breaks it into levels to aid in visualization, the underlying knowledge is conceived of as continuous. The levels differ primarily in terms of the types of equation structures with which children are successful. At Level 1, children succeed with the traditional operations-equals-answer structure (e.g., 8 + 4 = __) and view the equal sign as an operator signal to calculate an answer. At Level 2, children succeed on a wider variety of equation structures, including problems with operations on the right side of the equal sign (e.g., __ = 8 + 4) and problems with no operations (e.g., 3 = 3). At Level 3, children succeed on problems with operations on both sides of the equal sign and recognize a relational view of the equal sign as valid. Finally, at Level 4, children succeed on problems regardless of structure and generate a flexible, relational view of the equal sign. A hallmark of Level 4 understanding is recognizing transforma-

Table 1
Construct Map for Math Equivalence Knowledge

Level	Description	Equation structure		Sample items
4. Comparative relational	Successful with a variety of equation structures by comparing expressions on the two sides of the equal sign. Generate a relational definition of the equal sign.	Operations on both sides with multi-digit numbers or with multiple instances of a variable.	ST	67 + 86 = 68 + 85 (true or false? explain why)
			EQ	What does the equal sign mean?
			OE	__ + 55 = 37 + 54
3. Basic relational	Successful with equation structures with operations on both sides of the equal sign. Recognize a relational definition of the equal sign as correct.	Operations on both sides: $a + b = c + d$ $a + b - c = d + e$	ST	31 + 16 = 16 + 31 (true or false?)
			EQ	Is "the same as" a good definition of equal sign?
			OE	5 + __ = 6 + 2
2. Flexible operational	Successful with equation structures that are compatible with an operational view of the equal sign.	Operations on the right side: $c = a + b$ No operations: $a = a$	ST	4 = 4 + 0 (true or false?)
			EQ	10 cents __ one dime (select correct symbol)
			OE	7 = __ + 3
1. Rigid operational	Successful with an operations-equals-answer equation structure. Generate an operational definition of the equal sign.	Operations on the left side: $a + b = c$	ST	5 + 2 = 7 (true or false?)
			EQ	Identify a pair equal to 3 + 6
			OE	__ + 5 = 9

Note. Success includes the ability to solve, evaluate, and encode equations of a particular structure. ST = equation-structure items; EQ = Equal-sign items; OE = open equation-solving items. From "Assessing knowledge of mathematical equivalence: A construct modeling approach," by B. Rittle-Johnson, P. G. Matthews, R. S. Taylor, & K. McEldoon, K, 2011, *Journal of Educational Psychology, 103,* p. 87. Copyright 2011 by American Psychological Association. Adapted with permission.

tions that maintain the equality of an equation (e.g., adding the same number to both sides of the equal sign) without engaging in full computation strategies (e.g., Alibali et al., 2007; Carpenter et al., 2003; Jacobs et al., 2007; Steinberg et al., 1991).

In two previous studies, researchers documented the construction and validation of comprehensive assessments intended to measure math equivalence knowledge in terms of this construct map (Matthews et al., 2012; Rittle-Johnson et al., 2011). The assessment items varied in type (e.g., equation-solving, equation-structure, equal-sign) and in structural arrangement (e.g., operations-equals-answer, operations on both sides). The researchers used Rasch modeling procedures to measure item difficulties on a continuous scale. The result was two forms of an assessment that were reliable and valid. Further, the order of the empirically derived item difficulties supported the hypothesized construct map in two samples. Thus, this measurement work emerged from and was well aligned with the larger literature on math equivalence, and it resulted in a psychometrically sound assessment tool for elementary schoolchildren (ages 7 to 11).

The Current Study

The goal of the present study was to extend this work to an older cohort of middle school students. In the following text, we outline our three specific aims and the motivation for each.

Our first aim was to investigate whether the construct map and associated assessments could reliably measure knowledge of math equivalence among an older cohort of students (ages 12 to 16). Previous research indicates that difficulties with math equivalence persist well past elementary school (e.g., Alibali et al., 2007; Jones, Inglis, Gilmore, & Dowens, 2012; Knuth et al., 2006; Li et al., 2008; McNeil et al., 2006; Renwick, 1932). For example, Knuth et al. (2005) found that nearly half of the middle school students in the study provided an operational definition of the equal sign. Similarly, Booth and Davenport (2013) found the average percent correct on a measure of equal sign understanding was close to 50% for a sample of middle school students. Alibali et al. (2007) also demonstrated that fewer than 60% of students at the end of eighth grade provided a relational definition of the equal sign. More importantly, middle school students, high school students and adults sometimes exhibit operational views of equations even after they are taught about the equal sign or equivalence more generally (e.g., Herscovics & Kieran, 1980; Sfard & Linchevski, 1994). For example, educated adults still sometimes solve standard equivalence problems (e.g., 6 + 8 + 4 = 7 + __) using operational strategies, giving the answers 18 or 25 – particularly under speeded conditions (e.g., Chesney, McNeil, Brockmole, & Kelley, 2013; McNeil & Alibali, 2005; McNeil et al., 2010).

This body of research suggests that valid measures of math equivalence knowledge that function beyond elementary school are clearly needed. However, no studies to date have used a discriminating assessment based on a construct map to unpack the structure of math equivalence knowledge in students beyond age 11. We sought to fill this gap by validating the construct map and associated assessments in an older sample of middle school students.

Our second aim was to report quantitative and qualitative differences in math equivalence knowledge between students in pre-algebra and algebra classes. Many studies have considered middle school students as a homogenous population and have not focused on differences as a function of experience. However, a formal understanding of math equivalence is widely regarded as a necessary component of success in algebra (e.g., Carpenter et al., 2003; Jacobs et al., 2007; Knuth et al., 2006; NCTM, 2000; Steinberg et al., 1991), and students in formal algebra courses are more likely to have experiences that explicitly attempt to bridge the "cognitive gap" between arithmetic and algebra. That is, they are more likely to have experiences that rely on explicit recognition of the arithmetic properties of algebra (e.g., performing the same operation on both sides of the equation). A key question is whether these experiences in algebra support shifts in students' understanding of equivalence. We expect they do and that students in algebra classes will provide more relational responses on the equivalence assessment than pre-algebra students.

Our third aim was to examine the relation between knowledge of math equivalence and students' formal reasoning about algebraic expressions. Prior work suggests that knowledge of math equivalence is related to performance on algebraic equations with variables (e.g., Alibali et al., 2007; Booth & Davenport, 2013; Knuth et al., 2005). For example, middle school students who provided a relational definition of the equal sign were almost twice as likely to solve algebra equations correctly (e.g., $4m + 10 = 70$) than students who did not (Knuth et al., 2006). But, if math equivalence is truly foundational for algebraic thinking, it should predict performance on a variety of tasks—including tasks with algebraic expressions that do not include the equal sign.

Again, the notion of a "cognitive gap" between arithmetic and algebra is relevant (Linchevski & Herscovics, 1996. As noted, experiences with arithmetic can give rise to misconceptions about the equal sign and equivalence more generally (e.g., the idea that the equal sign is unidirectional and operational). In a similar way, experiences with arithmetic can also give rise to misconceptions about variables and expressions with variables (e.g., Kinzel, 1999; Lucariello et al., 2014; MacGregor & Stacey, 1997; McNeil, Weinberg, et al., 2010). For example, consider a 12-year-olds' difficulty assigning meaning to expressions such as $3a$, $a + 3$, and $3a + 5a$ because, "there is no equal sign with a number after it" (Kieran, 1981, p. 324), or consider a 13-year-olds' difficulty subtracting $8x$ from both sides of an equation because "I don't know how much is $8x$" (Sfard & Linchevski, 1994, p. 107). Both cases exemplify an operational viewpoint in which students treat expressions in terms of processes—signals to fill in a missing value—rather than objects that can be manipulated. Thus, in broader terms, students' concept of equation and all that it encompasses progresses from an operational view to a relational or structural view (Kieran, 1981; Sfard & Linchevski, 1994), and depending on where students are in this development, we would expect similar kinds of errors in thinking across both the math equivalence assessment and their reasoning about algebraic expressions.

Method

Participants

Participants represented a convenience sample of 229 students (106 female) from two public schools in a small city in the American West. One school served Grades 7 through 9 ($n = 165$) and had a student population that was 50% Caucasian, 48% Hispanic, and 86% qualified for free or reduced-price lunch. The other school ($n = 64$) served Grades 7 through 12 and had a student population that was 81% Caucasian, 13% Hispanic, 3% African American, and 29% qualified for free or reduced-price lunch. All students from six different teachers were invited to participate with no exclusion criteria. The majority of participants in this study (94%) were seventh- and eighth-grade students between 12 and 14 years old (M age = 13.1 years, SD = 0.8, minimum = 12.0, maximum = 16.0). Because one school spanned Grades 7 through 12, some students in the sample were in higher grades and somewhat older than the typical middle school sample. The majority of students (62%) were in pre-algebra classes. The remaining students were in Algebra I (34%) or an advanced secondary mathematics course (3%) that included a mix of algebra and geometry. Because our focus was on class experience (i.e., pre-algebra vs. algebra) rather than grade level, we refer to all students in our sample as "middle school" students. For conciseness, we also refer to both Algebra I students and advanced secondary math students as "algebra" students as these courses included algebra instruction. No records were collected regarding whether students required an individualize education plan or testing accommodations.

Materials

Math equivalence assessment. We administered two forms of the assessment from Matthews et al. (2012; see also Rittle-Johnson et al., 2011) with a few changes detailed below in the following text. Note that Rittle-Johnson et al. (2011) originally constructed two different, but comparable forms in order to facilitate pre-post testing in intervention studies. We continued using two comparable forms in the current design instead of settling on one for similar practical purposes.

There were three problem types. *Open-equation-solving* items assessed students' abilities to solve equations of varying difficulty. For example, an easy item included an operation on the right side of the equal sign (e.g., $8 = 6 + \_\_$), and more difficult items included operations on both sides of the equal sign (e.g., $7 + 6 + 4 = 7 + \_\_$). Several open-equation-solving items also included letter variables (e.g., $c + c + 4 = 16$). *Equation-structure* items assessed students' understanding of valid equation structures as well as their abilities to reason about both sides of an equation without computation. For example, one easier item had students evaluate whether $31 + 16 = 16 + 31$ was true or false, and a more difficult item had students explain why $67 + 86 = 68 + 85$ was true without adding the numbers together. *Equal-sign* items assessed students' explicit understanding of the symbolic equal sign. A core item asked students to define the equal sign, and another item had students rate different definitions of the equal sign as good or not good. A full list of the items used on both forms is included in the Appendix.

Each form of the assessment had 31 items (12 open-equation-solving, 13 equation-structure, and 6 equal-sign). Each item was classified a priori as tapping knowledge at Levels 1, 2, 3, or 4 of the construct map (see Table 1). We began with the forms from Matthews et al. (2012), which used a step-by-step item matching procedure to ensure similarity of content and difficulty across forms. However, we introduced a change to ensure that all parameters could be easily placed on the same measurement scale. One form of the assessment we used was identical to Form 1 from Matthews et al. (2012). The other was nearly identical to Form 2, with the key exception that we replaced a number of Form 2 items with those from Form 1 so that we would have 10 anchor items (denoted with the superscript "a" in the Appendix). We chose anchor items to span the range of item difficulties specified a priori by the construct map and distributed across different item types. We oversampled items hypothesized to be more difficult because we expected pre-algebra and algebra students to have higher ability than past samples, which only included elementary schoolchildren. The result was that 10 of the 31 items on each assessment were identical across forms, serving as anchors to aid in equating scores across forms. We applied a concurrent calibration procedure (Kolen & Brennan, 2004) to yield item and person parameters that were on a common metric. Due to experimenter error, one item on Form 1 was misprinted. It was originally intended to be an open-equation-solving item at Level 3 with operations on both sides of the equal sign. However, it was misprinted as __ = 8 + 5 + 9. Since this only contains operations on the right side of the equal sign, the misprint led us to designate it as a Level 2 item rather than a Level 3 item. This is marked in the Appendix (Form 1, Item 25).

Each item was scored dichotomously as correct (1) or incorrect (0). As in prior work, equation-solving answers within one of the correct answer were scored as correct to account for minor calculation errors (e.g., Perry, 1991; Rittle-Johnson, 2006). Ten items required students to provide a written definition or explanation, and responses were coded based on the system from Matthews et al. (2012). Specifically, responses were coded as correct if the student mentioned the equality relation between values on the two sides of the equal sign. For example, for defining the equal sign, responses of "it means the same as" or "the same amount" were coded as correct and responses of "the answer" or "the sum" were not. A second rater independently scored responses for 35% of the sample, and interrater agreement was high, with an average agreement of .95 on both Form 1 (range = .89–1.00) and Form 2 (range = .88–1.00).

Algebraic expression. In addition to the equivalence assessments, we administered an item that has been used to assess conceptual understanding of algebraic expressions (see McNeil, Weinberg, et al., 2010, see also Küchemann, 1978): "Cakes cost *c* dollars each and brownies cost *b* dollars each. Suppose I buy 4 cakes and 3 brownies. What does $4c + 3b$ stand for?" For students who completed Form 1 of the assessment, the specific symbols used were mnemonic in that the price of a cake was represented by *c* and the price of a brownie was represented by *b*. For students who completed Form 2 of the assessment, the specific symbols used were traditional in that the *c* and *b* were replaced by the traditional letters *x* and *y*. This contrast has been of interest to researchers because use of mnemonic symbols may strengthen students' naïve conceptions that variables in algebraic expressions stand for labels instead of quantities (McNeil et al., 2010; see also Küchemann, 1978; MacGregor & Stacey, 1997; Rosnick, 1981).

This problem differed in key ways from the items on the assessment that contained letter variables (i.e., $10 = z + 6$, $c + c + 4 = 16$, $m + m + m = m + 12$ on Form 1). First, the problems on the assessment contained the equal sign, but the algebraic expression did not. Second, the problems on the assessment required students to solve for the value of the variable, but the algebraic expression required students to conceptually interpret the symbols without any calculation. Third, none of the problems on the assessment featured products symbolized by the juxtaposition of a variable and a coefficient (e.g., $3x$), which is a more difficult symbolic form to understand relative to a stand-alone variable. Thus, this problem tapped students' formal understanding of symbolic letter variables and their interpretation within an algebraic expression.

Table 2
Examples of Students' Interpretations of the Algebraic Expression

Interpretations of the expression $4c + 3b$
Correct interpretations
"The total cost of cakes and brownies"
"The amount of money you'll pay for the cakes and brownies"
"The price of four cakes plus the price of three brownies"
"Four times the cost of one cake plus three times the cost of one brownie"
Incorrect interpretations
"Four cakes plus three brownies"
"The number of cakes and brownies I bought"
"It means you do 4 times *c* and 3 times *b*"
"That is the equation to find the answer"

Responses were coded based on a system developed in previous research (McNeil, Weinberg, et al., 2010). A response was scored as correct (1) if the student indicated that the letters stood for the cost or price of the cakes and brownies (see Table 2 for examples). A second rater scored responses for 35% of the sample, and interrater agreement was high for coding responses as correct or incorrect (.95).

Procedure

Assessments were administered to whole classes using an alternating procedure (i.e., we alternated handing out Form 1 and Form 2 so the first student got Form 1, the second student got Form 2, the third student got Form 1, the fourth student got Form 2, etc.). This ensured that about the same number of each form was distributed in each class ($n_{V1} = 114$, $n_{V2} = 115$) and that the distribution would result in randomly equivalent groups taking the two forms (Kolen & Brennan, 2004). The algebraic expression problem was printed as the final problem on a separate page of the assessment. The sessions lasted about 45 min. The groups assigned to different test forms were similar in terms of mean age (Form 1 = 13.0, Form 2 = 13.1), percent female (Form 1 = 46%, Form 2 = 47%), and percent in pre-algebra (Form 1 = 63%, Form 2 = 62%).

Data Analysis

We used a Rasch model to examine performance on the assessment. Rasch modeling is a one-parameter member of the item response theory (IRT) family (Bond & Fox, 2007). The Rasch model estimates item difficulty and student ability levels simultaneously, yielding the probability that a particular respondent will answer a particular item correctly (Rasch, 1960/1993; Wright, 1977). We used Winsteps software (3.80.1; Linacre, 2013) to perform all IRT estimation procedures using default settings. Given our use of a common-item design, we applied a concurrent calibration procedure to ensure that the parameter estimates for each form were calibrated to the same scale (e.g., Kolen & Brennan, 2004). Because this approach assumes the common items function equivalently across groups (i.e., invariance), we conducted a check on this assumption by performing separate calibrations of each form and examining the relationship between difficulty estimates for common items across forms. The presence of invariance would be supported by a strong linear relationship between estimates. Specifically, we inspected a scatterplot and found the best fitting linear estimate comparing the estimates from Form 1 and Form 2. As can be seen in Figure 1, the best-fitting line

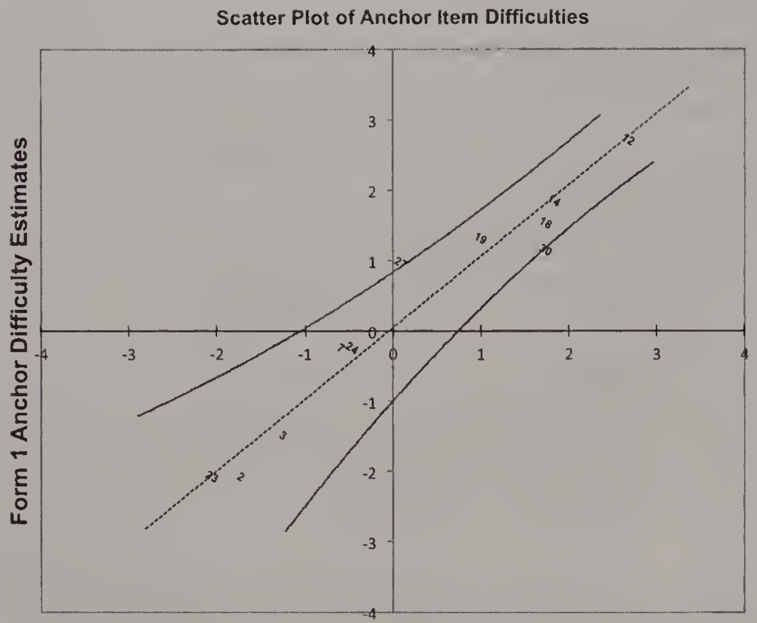


Figure 1. Scatterplot of anchor item difficulties for Form 1 versus Form 2. Plots anchor item difficulty estimates in logits for both forms of the assessment. Best fit line has a slope near 1 and an intercept near 0, indicating that the scales for each form are interchangeable.

had a slope of 1.02, an intercept of .05, and all items were close to the regression line. This supports the invariance assumption and suggests that concurrent estimation of all items from both forms is warranted (Kolen & Brennan, 2004; Linacre, 2016). Thus, all estimates discussed below are from the concurrent estimation (of 53 total items, if anchors are not double counted) and can be interpreted against a common scale.

Results

First, we examine the psychometric properties of the math equivalence assessment to determine whether it functions well for students in middle school. Second, we compare performance as a function of students' current math course (pre-algebra vs. algebra). Finally, we describe performance on the algebraic expression item, and test whether students' knowledge of math equivalence is correlated with their interpretation of the algebraic expression.

Math Equivalence Assessment

Rasch model fit information supported the unidimensionality of the assessment, indicating that it largely tapped a single construct. Unidimensionality in Rasch modeling is often assessed by principal components analysis (PCA; e.g., Bond & Fox, 2007; Hattie, 1985; Smith, 1996). Specifically, the model accounted for 45.6% of the variance in our data set (eigenvalue of 43.6). The largest secondary factor accounted for 2.6% of the variance (eigenvalue of 2.5). The Rasch model can also be evaluated using infit and outfit statistics, which indicate different types of problematic items (Bond & Fox, 2007). Infit statistics detect unexpected responses to items with difficulty estimates close to respondents' ability estimates. In contrast, outfit statistics tend to reflect the influence of unexpected responses to items that are far from respondents' ability estimates. All items on the assessment had good infit values within the range of 0.5 and 1.5 (Linacre, 2016; Wright & Linacre, 1994). By contrast, 19 items had outfit statistics outside the .5 to 1.5 range, and 17 of those items were easy items with near-ceiling performance ($M_{\text{accuracy}} = 96.2\%$, $SD = 2.6\%$). These outfit results were perhaps to be expected given that the assessment was originally intended for elementary-school-age children and, as such, included several lower-level items that 7th through 9th grade students should complete easily. We followed the advice of Bond and Fox (2007) and Linacre (2016) and focused more on infit than on outfit measures. See Table 3 for item difficulty estimates and fit measures.

The PCA results along with infit indices support the use of Rasch analysis. As a supplementary analysis, we also conducted two sets of confirmatory factor analysis using MPlus software (Muthén & Muthén, 1998-2017) for each assessment version: (a) a one-factor model that included all items as loading on a single factor and (b) a three-factor model that separated items according to each of the three question types (structure, equal sign, and solve), while allowing each of the three factors to covary with the others. The results are summarized in Table 4. For both versions of the assessment, RMSEA for the one-factor model was ≤.067 and essentially equivalent to the RMSEA for the three-factor model (as indicated by nearly identical confidence intervals). Moreover, the ratio of chi-square to degrees of freedom was relatively low for the 1-factor model. Although the CFI for both the one- and three-factor models was low compared to a desired benchmark of about .9, this was probably due in part to the lower level of statistical dependence observed across items and in part due to the large number of items on each form. The low CFI can be interpreted as a sign that the single underlying dimension is not particularly strong, although that does not preclude the unidimensionality assumption being appropriate (indeed, CFI is not high for the three-factor model either).

On balance, our analyses suggest that the unidimensionality assumption proved adequate with the caveat that a number of items demonstrate some degree of misfit—primarily indicated by outfit measures—because of items with low difficulty levels. This issue seems to be an inherent difficulty of using an assessment that includes measures for the lowest levels of our construct map with middle school students (e.g., solve _ + 5 = 9). Generally speaking, items with difficulties that are very low for the sample in question simply yield low information and can be prone to some sort of misfit. We briefly return to this issue when addressing limitations of our method in the Discussion section.

Beyond demonstrating adequate evidence for the unidimensionality assumption, the assessments were consistent and showed adequate capacity to resolve person ability and item difficulty estimates. Item reliability as assessed by the Rasch model was generally good ($R_I = .98$), indicting that sample size was large enough to estimate item difficulty well (Linacre, 2016). Person reliability was ($R_P = .78$) was just short of the normative cutoff of .80, indicating that although adequate, more items may be needed to adequately distinguish between high and low performers. Given that the measures were designed using a younger sample, it is noteworthy that the item and person reliabilities remained adequate with the current sample.

We also evaluated whether our hypothesized levels of difficulty matched the empirical estimates. Recall that we selected

Table 3
Item Statistics for Mathematical Equivalence Assessment

Construct component	Number	Hypothesized level	Item name	Item example	Mean accuracy	Observed item location (δ, in Logits)	Standard error of δ estimate	Infit MSQ	Outfit MSQ
Structure	1[a]	1	L1_Structure1_V1V2	$8 = 5 + 10$ (True or False)	.96	−1.57	.36	.9544	.6267
Structure	2	2	L2_Structure2_V1	$8 = 8$ (True or False)	.97	−2.07	.61	1.0946	.5365
Structure	3	2	L2_Structure3_V1	$4 = 4 + 0$ (True or False)	.96	−1.74	.54	1.0287	1.3742
Structure	4	2	L2_Structure4_V1	$8 = 5 + 3$ (Explain why True)	.97	−2.03	.62	.7506	.1427[b]
Structure	5[a]	3	L3_Structure5_V1V2	$31 + 16 = 16 + 31$ (True or False)	.96	−1.57	.36	.9006	.5156
Structure	6	3	L3_Structure6_V1	$3 + 1 = 1 + 1 + 2$ (True or False)	.95	−1.25	.45	1.0937	1.5785
Structure	7[a]	3	L3_Structure7_V1V2	$6 + 4 = 5 + 5$ (Explain why True)	.9	−.37	.25	.7184	.542
Structure	8	4	L4_Structure8_V1	$89 + 44 = 87 + 46$ (Explain why True)	.59	2.02	.22	1.0444	1.153
Structure	9	4	L4_Structure9_V1	$1 + 9 + \square = 10 + \square$ (Explain which numbers could go in box)	.51	2.45	.22	.9688	.8884
Structure	10	4	L4_Structure10_V1	If $25 + 14 = 39$, does $25 + 14 + 7 = 39 + 7$? (explain why)	.59	2.02	.22	.8435	.7927
Structure	11	4	L4_Structure11_V1	If $3 \times 5 = 15$, does $3 \times 5 \times 6 = 15 \times 6$? (explain why)	.49	2.55	.22	.9907	.9444
Structure	12[a]	4	L4_Structure12_V1V2	If $76 + 49 = 121$, does $76 + 49 - 9 = 121 - 9$? (explain why without subtracting)	.47	2.74	.16	.8756	.8002
Structure	13	4	L4_Structure13_V1	$3 \times \_\_ = 45, 7 \times 3 \times \_\_ = 7 \times 45$ (explain why the same number goes in the blanks)	.15	4.89	.31	.9575	.6901
Equal-sign	14	1	L1_EqualSign14_V1	$3 + 6$ (Identify an equal pair)	.92	−.73	.38	1.2629	2.6169
Equal-sign	15	2	L2_EqualSign15_V1	10 cents __ one dime (select symbol that shows they are the same)	.96	−1.74	.54	1.0759	.4256[b]
Equal-sign	16	3	L3_EqualSign16_V1	Is "two amounts are the same" a good definition of the equal sign?	.93	−.89	.4	.9976	.502
Equal-sign	17	4	L4_EqualSign17_V1	Which (of three options) is the best definition of the equal sign?	.67	1.56	.23	1.0724	1.1386
Equal-sign	18[a]	4	L4_EqualSign18_V1V2	1 quarter = 25 pennies (define equal sign in this context)	.73	1.2	.17	1.1439	1.0716
Equal-sign	19[a]	4	L4_EqualSign19_V1V2	What does the equal sign mean?	.62	1.88	.16	1.0508	1.0247
Solving	20	1	L1_Solve20_V1	$\_\_ + 5 = 9$	.98	−2.52	.74	1.1714	7.6052[b]
Solving	21	2	L2_Solve21_V1	$7 = \_\_ + 3$	.99	−3.26	1.02	.9759	.1575[b]
Solving	22[a]	3	L3_Solve22_V1V2	$3 + 6 = 8 + \_\_$	.97	−2.04	.43	.8617	1.5427[b]
Solving	23	3	L3_Solve23_V1	$5 + \_\_ = 6 + 2$	.96	−1.74	.54	.8749	.2063[b]
Solving	24[a]	3	L3_Solve24_V1V2	$7 + 6 + 4 = 7 + \_\_$	.9	−.31	.24	1.0839	1.9299[b]
Solving	25	2	L2_Solve25_V1	$\_\_ = 8 + 5 + 9$	.94	−1.06	.42	1.1763	5.1268[b]
Solving	26	3	L3_Solve26_V1	$8 + 5 - 3 = 8 + \_\_$	.9	−.46	.35	.9169	.9855
Solving	27[a]	4	L4_Solve27_V1V2	$67 + 84 = \_\_ + 83$	.81	.61	.19	.8555	.6237
Solving	28	4	L4_Solve28_V1	$\_\_ + 55 = 37 + 54$	.75	.98	.25	1.0703	1.9032[b]
Solving	29	3	L3_Solve29_V1	$13 = n + 5$	.92	−.73	.38	1.1461	1.0362
Solving	30[a]	4	L4_Solve30_V1V2	$c + c + 4 = 16$	.69	1.49	.17	1.018	1.1409
Solving	31	4	L4_Solve31_V1	$z + z + z = z + 8$	.97	1.77	.23	.9584	.8911

Note. The data table is based on collapsing the data from the two forms of the revised assessment, with example items from one of the forms. L = level; V = version.
[a] Indicates anchor items. [b] Indicates items with outfit scores out of the desired range. These items tended to be items for which the current sample demonstrated ceiling effects.

items to tap four different levels of knowledge, as outlined in the construct map (see Table 1). The hypothesized level of difficulty for each item (1, 2, 3, or 4; see the Appendix) correlated highly with the empirically derived item difficulty estimates, Spearman's $\rho(52) = .88, p < .001$. We further used a Wright map (Wilson, 2005) to visually inspect the difficulty of the items (see Figure 2). A Wright map has two columns, one for respondents and one for items. Respondents with higher ability estimates are near the top of the map and those with lower ability estimates are near the bottom. Similarly, items with higher difficulty estimates are near the top of the map and items with lower difficulty estimates are near the bottom. The vertical line indicates the scale for the ability and difficulty estimates measured in logits (i.e., log-odds unit). Average difficulty was set to 0 logits. We added horizontal lines to visually highlight the clustering of items by hypothesized level. However, it should be explicitly noted that the construct is a continuous measure and that the link between student ability estimates and item difficulty estimates is a probabilistic one. Thus, the lines we added for clarity should not be interpreted as discrete stages.

As shown on the Wright map and in Table 3, the items we had a priori categorized as Level 4 items proved to be the most difficult (i.e., clustered near the top of the Wright map). All of the Level 4 items had higher difficulty estimates than any of the Level 1, 2, or 3 items. The items we a priori categorized as Levels 1, 2, and 3 were somewhat less distinct, but tended to cluster in the expected order. Most Level 3 items had difficulty estimates near −1, Level 2 items had difficulty estimates near −2, and Level 1 items had difficult estimates near −2.5.

There was a small set of items (9 out of 62, 14.5%) that did not function according to our hypothesized levels. Five were open-equation-solving items that were easier than expected. Two Level 2 items ($8 = 6 + \_\_$, $7 = \_\_ + 3$) functioned more like Level 1

Respondents	Logits	Items
xxxxx	6	
xxxxxxxxxx	|	
	5.5	
xxxxxxxx	|	
xxxxx	5	L4_Structure13_V1; L4_Structure13_V2
	|	
	4.5	
xxxxxxxxxxx	|	
xxxxxxxxxxxx	4	
xxxxxxxxx	|	
xxxxxxxxxxxxxx	3.5	
xxxxxxxxxxxxxx	|	
xxxxxxxxxx	3	L4_Structure11_V2
xxxxxxxxxxxxxxxxxxxx	|	L4_Structure11_V1; L4_Structure12_V1V2; L4_Structure10_V2
xxxxxxxxxxxx	2.5	L4_Structure9_V2
xxxxxxxxxxxxxxxxxxxxxxxx	|	L4_Structure9_V1; L4_Solve31_V2
xxxxxxxxxxxxxx	2	L4_Structure10_V1; L4_Structure8_V1; L4_Structure8_V2
xxxxxxxxxxxxxxx	|	L4_EqualSign17_V2; L4_Solve31_V1; L4_EqualSign_19_V1V2
xxxxxxxxxxxxx	1.5	L4_EqualSign17_V1; L4_Solve30_V1V2
xxxxxxxxx	|	L4_EqualSign18_V1V2
xxxxxxx	1	L4_Solve28_V1; L4_Solve28_V2
xxxxxxxx	|	
x	0.5	L4_Solve27_V1V2
xxx	|	
x	0	
xxxx	|	L3_Solve24_V1V2; L3_Structure7_V1V2; L3_Structure6_V2
xxxxx	-0.5	L3_Solve26_V1; L3_EqualSign16_V2
xxx	|	L3_Solve25_V2; L3_Solve29_V1; L1_EqualSign14_V1
	-1	L3_Solve26_V2; L2_Structure3_V2; L3_EqualSign_16_V1
x	|	L2_Solve25_V1; L3_Structure6_V1
	-1.5	L3_Solve29_V2
x	|	L3_Structure5_V1V2; L1_Structure1_V1V2; L2_Structure2_V2
	-2	L2_Structure3_V1; L2_EqualSign15_V1 L3_Solve23_V1; L22_Structure4_V2
	|	L2_Structure4_V1; L2_Structure2_V1; L3_Solve22_V2; L3_Solve23_V1
	-2.5	L1_Solve20_V2; L1_EqualSign14_V2; L2_EqualSign15_V2
	|	L1_Solve20_V1
	-3	L2_Solve21_V1; L2_Solve21_V2

Figure 2. Wright map for the math equivalence assessment. Each "x" on the left represents one student. Each entry on the right represents one item. Item entries name the hypothesized difficulty level (e.g., L4), the item type (e.g., structure), the item number (e.g., 13), and the form number (e.g., V1). The numbers on the vertical axis represent item difficulty and student ability estimates in logits. The horizontal lines are for visual, descriptive purposes only—the construct is theorized to be continuous.

items. This occurred in prior work as well (Matthews et al., 2012; Rittle-Johnson et al., 2011), suggesting we need to reevaluate where these items fall on the construct map. Three Level 3 items (3 + 6 = 8 + __, __ + 2 = 6 + 4, 5 + __ = 6 + 2) functioned more like Level 2 items, which could be due to error/noise or to these items functioning differently among this older, more experienced sample.

There were four items that were harder than expected. Three were Level 1 items (pick a pair equal to 6 + 3 and judge 8 = 5 + 10 as true/false on both forms). However, these items' difficulty estimates were rank ordered similarly to that reported in previous research (Matthews et al., 2012; Rittle-Johnson et al., 2011). The last was a Level 2 item, the misprinted solve item on Form 1 (__ = 8 + 5 + 9), perhaps more difficult because it contained three addends instead of two. However, only four students missed this item suggesting it was still relatively easy.

Overall, the assessment performed well, and the construct map seemed to apply to this older sample. Specifically, different item types with varying levels of difficulty were measured on a single scale and functioned in a way that matched the hypothesized construct map. We acknowledge a caveat to this point: The sample studied here is different from the sample in earlier studies (Matthews et al., 2012; Rittle-Johnson et al., 2011), so we cannot directly compare item difficulty estimates. Thus, to evaluate similarity of the construct map's performance and applicability across studies with different age groups, we relied on the similarity of the rank orderings (i.e., rankings from low difficulty to high difficulty estimates) across samples.

Table 4
Confirmatory Factor Analysis Exploring Unidimensionality

	Version 1		Version 2	
Measure	1-Factor Model	3-Factor Model	1-Factor Model	3-Factor Model
RMSEA	.067	.062	.045	.043
RMSEA CI	(.056, .077)	(.051, .073)	(.038, .053)	(.035, .051)
CFI	.725	.762	.762	.735
χ^2	652.831	619.710	637.467	613.687
df	434	432	434	321

Note. RMSEA = root mean square error of approximation; CI = confidence interval; CFI = comparative fit index.

Math Equivalence Performance as a Function of Math Course

Students performed well on the assessment ($M_{\text{accuracy}} = 79\%$ [25 out of 31], $SD = 14\%$), but only 2% of the sample scored at ceiling. An analysis based on percent correct revealed differences as a function of math course after controlling for age. Specifically, pre-algebra students scored significantly lower ($M = 77\%$, $SE = 1\%$) than algebra students ($M = 83\%$, $SE = 2\%$), $F(1, 226) = 4.45, p = .04, \eta_p^2 = .02$. Analyses based on Rasch ability estimates supported these results. As shown on the Wright map, there was an approximately normal distribution of ability estimates. Further, ability estimates ($M = 2.57$, $SD = 1.50$, range $= -1.56$ to 6.66) were positively correlated with students' self-reported expected grade in the class (A, B, C, D, or F), Spearman's $\rho(229) = .41, p < .001$. Ability estimates also differed significantly by math course after controlling for age, $F(1, 226) = 4.95, p = .03, \eta_p^2 = .02$, such that pre-algebra students had lower ability estimates ($M = 2.3$, $SE = 0.1$) than algebra students ($M = 2.9$, $SE = 0.2$).

To further examine this group difference, we looked at the probability of success on items of different difficulty levels for students at different ability levels. The model allows us to calculate the probability of any participant's success on any given item from log-odd units by using the following equation:

$$Pr(success) = \frac{1}{1 + e^{-(\theta - d)}},$$

in which θ is a participant's ability estimate and *d* is the item difficulty estimate. First, we selected six items from Table 3: two items with the lowest and highest difficulty estimates and four items with difficulty estimates that represented each of the four levels on the construct map (e.g., items we a priori categorized as Level 1 had difficulty estimates near −2.5, so one item we selected had an empirically derived difficulty estimate of −2.52). Second, we calculated the mean ability estimates for the pre-algebra group and the algebra group. Table 5 presents the probability of success for items at different difficulty levels for students at the mean ability level for each group. Importantly, the model predicts substantial differences in performance on typical Level 4 items as a function of math course, but predicts few differences for lower level items, as the typical student in both groups is expected to have high probabilities of success.

We qualitatively explored performance on three Level 4 items on which the differences between pre-algebra students and algebra students were particularly pronounced. The first item was a structure item: "17 + 12 = 29 is true. Without adding the 8, can you tell if 17 + 12 + 8 = 29 + 8 is true or false? How do you know?" Only 39% of pre-algebra students responded correctly compared to 64% of algebra students. A common correct response was to write, "You added the same amount to both sides so it's still equal." Students' incorrect responses revealed key differences. Of the pre-algebra students who answered incorrectly, 47% selected false or do not know (as opposed to true) indicating a conceptual misunderstanding. Their false/do not know selections were accompanied by explanations underscoring the fact that they thought adding the eights made the problem unequal (e.g., "If you add 8, it won't equal the same" or "You're adding 8 so the answer will go

Table 5
Probabilities of Success Based on Item Difficulty Estimates and Student Ability Estimates

			Probability of success	
Item	Hypothesized difficulty level	Difficulty estimate, δ	Pre-algebra mean ability estimate, θ = 2.3	Algebra mean ability estimate, θ = 2.9
7 = __ + 3	2	−3.26	.996	.998
__ + 5 = 9	1	−2.52	.992	.996
4 = 4 + 0 (true or false)	2	−1.74	.983	.990
13 = *n* + 5	3	−.73	.954	.974
If 25 + 14 = 39, does 25 + 14 + 7 = 39 + 7? (explain why)	4	2.02	.574	.711
3 × __ = 45, 7 × 3 × __ = 7 × 45 (Explain why the same number goes in the blanks)	4	4.89	.069	.119

Note. Entries in the two right columns represent the probabilities that the average student of a given ability estimate (2.3 or 2.9) will answer an item of a given difficulty estimate correctly. Difficulty level is the hypothesized level based on the construct map. Difficulty estimate is the empirically-derived difficulty estimates based on the Rasch analysis.

up"). In contrast, of the algebra students who answered incorrectly, only 30% were wrong because they selected false or do not know. The other 70% indicated that the equation was still true, but they had difficulty explaining their selection without calculating each side (e.g., "both sides equal 37"). Thus, not only were algebra students more likely to solve the problem correctly than pre-algebra students, even their errors were more indicative of an emerging relational understanding of equivalence.

The second item was an equal-sign item: "What does the equal sign mean in the statement: 1 quarter = 25 pennies?" Sixty-nine percent of pre-algebra students defined the equal sign relationally compared to 86% of algebra students. Common relational responses were to write, "the same as" or "the same amount of money." The majority of nonrelational responses did not necessarily reflect misconceptions, but were insufficient to convey that equality was understood rather than simply parroted back (e.g., "equal" "equal to" "they are equal"). However, some nonrelational responses reflected a common, operational misconception of the equal sign, often prevalent among elementary school students (e.g., "it means the answer to the problem" "the total"). Indeed, of all the pre-algebra students' incorrect responses, 17% of them were operational, compared to only 6% of algebra students' incorrect responses.

The third item with pronounced differences between groups was an open-equation-solving item: "Solve for c in the following equation, $c + c + 4 = 16$." Fifty-nine percent of pre-algebra students solved this item correctly compared to 73% of algebra students. Most often, students did not show work and simply wrote 6 as their answer. However, some correct responses were accompanied by written work and indicated a formal algebraic strategy (e.g., combining like terms to get $2c + 4 = 16$, subtracting 4 on both sides, and dividing both sides by 2). Twelve percent of pre-algebra students' correct responses were solved using this algebraic strategy compared to 22% of algebra students' correct responses. Incorrect responses were varied, but fell into one of four categories: blank/do not know (33% of incorrect responses), answering 4 (30%), answering 12 (20%), or other (17%). The proportion of each type of incorrect response was similar for pre-algebra and algebra students.

Overall, these results indicate that pre-algebra and algebra students did well on the math equivalence assessment. However, as expected, pre-algebra students had lower ability estimates than algebra students, and the Level 4 items were particularly key for capturing differences.

Interpreting Algebraic Expressions

Overall, 52% of the students interpreted the algebraic expression correctly by indicating that the variables stood for the costs of the cakes and brownies. We used logistic regression to examine whether the likelihood of interpreting the expression correctly depended on the type of variable used (x-and-y vs. c-and-b) and current math course (pre-algebra vs. algebra). We used Odds Ratios (OR) as our measure of effect size. For students in pre-algebra, those in the x-and-y condition were significantly more likely to interpret the expression correctly than those in the c-and-b condition (59% vs. 36%; $B = 0.94$, $SE = 0.34$, $p = .006$, $OR = 2.56$). For students in algebra, those in the x-and-y condition and those in the c-and-b condition were equally likely to interpret the expression correctly (55% vs. 66%; $B = -0.51$, $SE = 0.45$, $p = .25$, $OR = 0.60$). This difference was reflected by a significant variable type by math course interaction ($B = 1.45$, $SE = 0.56$, $p = .01$, $OR = 4.27$). Thus, for pre-algebra students, the use of mnemonic letters interfered with their ability to conceptually interpret the expression, whereas algebra students exhibited a deeper conceptual understanding of the expression that was not influenced by the specific letter variables used.

We also descriptively examined students' errors. The most common error was to use the letters as labels for the objects (e.g., three cakes and four brownies) rather than as the cost of the objects, accounting for 41% of all errors. Other errors included writing a literal translation of the expression (e.g., "3 multiplied by x plus 4 multiplied by y"; 23% of errors), responding in vague, uninterpretable ways (e.g., "it is the equation"; 20% of errors), adding unlike terms (e.g., "it must be $7xy$"; 9% of errors), or stating an inability to solve the problem (e.g., "I don't know"; 6% of errors). The letters as labels error was more common for students in the c-and-b condition (46% of errors) than for students in the x-and-y condition (35% of errors).

Finally, we tested whether students' knowledge of math equivalence was predictive of their interpretation of the algebraic expression. We used logistic regression to examine whether the likelihood of interpreting the expression correctly depended on empirically derived estimates of student ability on the math equivalence assessment. We included ability estimates as the primary predictor, as well as math course, assessment form, and students' age as control variables. Ability estimates were significantly predictive of success interpreting the algebraic expression ($B = 0.31$, $SE = 0.10$, $p = .001$, $OR = 1.37$). The remaining predictors were not significant when controlling for the others: math course ($B = 0.53$, $SE = 0.38$, $p = .16$), assessment form ($B = 0.39$, $SE = 0.28$, $p = .16$), and age ($B = -0.07$, $SE = 0.21$, $p = .75$). These results were consistent with our hypothesis that knowledge of math equivalence would be related to students' conceptual understanding of algebraic variables.

Recall that several items on the math equivalence assessment contained literal variables (e.g., "Solve for c in $c + c + 4 = 16$"). To ensure that the association between math equivalence knowledge and interpretation of the algebraic expression did not depend on the items involving literal variables, we ran a secondary analysis. Specifically, we obtained empirically derived ability estimates on the math equivalence assessment items that remained after excluding the five items with variables (out of 52). Ability estimates were still significantly predictive of success interpreting the algebraic expression ($B = 0.30$, $SE = 0.10$, $p = .002$, $OR = 1.35$), and the remaining predictors were not significant when controlling for the others: math course ($B = 0.54$, $SE = 0.38$, $p = .16$), assessment form ($B = 0.34$, $SE = 0.28$, $p = .22$), and age ($B = -0.07$, $SE = 0.21$, $p = .75$). Thus, students' knowledge of math equivalence—even when assessed without items using literal variables—was related to their conceptual interpretation of algebraic notation.

Discussion

Math equivalence is considered a "big idea" in mathematics as it lays a foundation for algebraic reasoning and for understanding math more generally (Charles, 2005; Jacobs et al., 2007; Kieran, 1981; Knuth et al., 2006; MacGregor & Stacey, 1997; NCTM, 2000). Thus, measuring knowledge of math equivalence is of clear importance. The current study extended the construct-modeling approach to measuring symbolic equivalence knowledge in three ways. First, we demonstrated that the equivalence assessment and construct map applied beyond elementary school, performing well with an older cohort of algebra and pre-algebra students. Second, we documented differences in math equivalence knowledge between students in pre-algebra and algebra classes, which were primarily captured by the difficult Level 4 items. Third, we confirmed that students' understanding of math equivalence was related to their interpretation of an algebraic expression, even after controlling for age and math course and after excluding the items containing variables on the math equivalence assessment. Below we outline the theoretical, practical, and methodological contributions of this research as well as potential future directions.

The results of the current study verified the validity of the math equivalence construct map explicated by Rittle-Johnson and colleagues (2011; see also Matthews et al., 2012) within an older and more mathematically experienced sample of middle school students. The items on the math equivalence assessment functioned according to the construct map, with key factors predicting item difficulty as hypothesized. This suggests that the difficulty of the equivalence construct has a stable order, supporting a key assumption of the Rasch model (Rasch, 1960/1993; Wright, 1977). Further, this is a good indication of a generalizable assessment that can be used vertically, at least from early elementary school to middle school algebra. This has important practical applications because previous work demonstrates that middle school students continue to struggle with understanding math equivalence (Alibali et al., 2007; Jones, Inglis, Gilmore, & Dowens, 2012; Knuth et al., 2006; Li et al., 2008; McNeil et al., 2006; Renwick, 1932) thus highlighting the need for assessments that can both track students' formal knowledge of math equivalence and serve as valid and reliable outcomes measures for intervention work.

As in prior work, the ordering of the item difficulties on the math equivalence assessment confirms that the structure of an equation is a key indicator of complexity and is therefore likely to influence performance (e.g., Baroody & Ginsburg, 1983; Matthews et al., 2012; Powell, Kearns, & Driver, 2016; Rittle-Johnson et al., 2011; Weaver, 1973). Specifically, the greater the structure deviates from the standard operations-equals-answer structure, the more difficult the problem is likely to be. This is true regardless of the specific task. For example, Figure 2 shows that open-equation-solving items are not inherently more difficult than equal-sign-definition problems (or vice versa). Rather, the difficulty depends on the structure of the equation and the extent to which the required solution strategy demands engaging arithmetic principles of equivalence as opposed to simple calculation. This has potential practical implications for designing interventions focused on varying problem structures, rather than varying problem tasks per se. Indeed, this is consistent with intervention research that has facilitated understanding of math equivalence by including practice with nonstandard equation structures (e.g., $17 = 9 + 8$; McNeil, Fyfe, & Dunwiddie, 2015; McNeil, Fyfe, Petersen, Dunwiddie, & Brletic-Shipley, 2011) or instruction on the meaning of the equal sign in the context of nonstandard equation structures (e.g., Fyfe & Rittle-Johnson, 2016; Fyfe, DeCaro, & Rittle-Johnson, 2014; Matthews & Rittle-Johnson, 2009; Perry, 1991; Powell & Fuchs, 2010).

In addition to validating the construct map, the assessment had considerable resolving power to detect variability in student knowledge. Even though students did well overall, there were reliable knowledge differences between students in pre-algebra and algebra. In particular, the model predicted substantial differences in performance on typical Level 4 items. Students in algebra were in fact more likely to exhibit comparative relational understanding by reasoning about transformations that preserve equality without reverting to calculation (e.g., "if we know $17 + 12 = 29$, can we tell if $17 + 12 + 8 = 29 + 8$ is true without adding?"; see Alibali et al., 2007; Matthews et al., 2012; Steinberg et al., 1991). These Level 4 items highlight that subtle differences not tapped by more typically used math equivalence items remain important for assessing students' knowledge of equivalence. Indeed, with the exception of defining the equal sign, the majority of past research has focused on items that tap understanding at Levels 1, 2, and 3 of the construct map (e.g., Alibali, 1999; Baroody & Ginsburg, 1983; Li et al., 2008; McNeil & Alibali, 2005). From an item response theory perspective, the Level 4 items add important information about learners who have moved beyond the basic levels of equivalence knowledge. From a practical perspective, this suggests Level 4 items should be included in assessments of math equivalence knowledge in order to obtain a nuanced picture of student understanding.

The inclusion of Level 4 items also provided empirical evidence for a solid connection between knowledge of symbolic math equivalence and at least some aspects of formal algebra (e.g., Alibali et al., 2007; Knuth et al., 2006; MacGregor & Stacey, 1997; Steinberg et al., 1991). Several Level 4 equation-solving items were basic algebra problems with an unknown variable (e.g., $c + c + 4 = 16$). As in prior work (Matthews et al., 2012), these items loaded highly on the math equivalence construct. That is, they fit well with the other items and functioned in predictable ways, even in this cross-section of pre-algebra and algebra students. This provides evidence that developing knowledge of algebra is strongly linked to knowledge of equivalence.

We also generated new findings about the links between equivalence knowledge and interpretation of variables: knowledge of math equivalence was related to students' conceptual interpretations of an algebraic expression that did not explicitly contain the equal sign. For pre-algebra students, the use of mnemonic letters (c and b to stand for the cost of cakes and brownies as opposed to the more traditional x and y) interfered with their ability to conceptually interpret the expression (see McNeil et al., 2010). In contrast, algebra students exhibited a deeper understanding of the expression that was not influenced by the specific letter variables used. Importantly, students' empirically derived ability estimates on the math equivalence assessment predicted their likelihood of interpreting the algebraic expression correctly, even after controlling for their current math course, the letter variables used, and their age. This lends support to the notion that a nuanced understanding of math equivalence extends to the concept as a whole beyond the use of the formal "=" symbol. It supports the broader idea that students' conception of math equivalence progresses from an operational view to a relational/structural view (e.g., Kieran, 1981; Sfard & Linchevski, 1994) and that where they are in this progression predicts their reasoning about expressions with variables on a formal algebra task.

A final contribution of the current research is to reinforce the benefits of combining quantitative and qualitative methodological approaches in integrative ways. For example, many of the Level 4 items required a qualitative coding of students' written responses. It was often insufficient to know whether the student judged the equation as true or false. Rather, we had to take into account the student's written explanation and to look for evidence of comparative relational understanding. These coding schemes were influenced heavily by qualitative work in mathematics education, such as that by Behr et al. (1980) and Carpenter et al. (2003). After the initial qualitative coding of student responses, a quantitative psychometric approach was applied (i.e., the Rasch model), which allowed us to obtain empirical estimates of item difficulties and student abilities. Finally, differences in item difficulties and student ability estimates helped to identify the items on which students varied in key ways, allowing us to take a closer, qualitative look at students' errors on those items. Thus, this iterative process not only showed that the qualitative and quantitative analyses were fully compatible, but also provided greater insight into the structure of students' knowledge than either approach alone.

Despite the contributions of the current research, there are a number of limitations. First, although we provided some evidence for the validity of the assessment, we did not include additional measures that would allow us to assess discriminant validity (e.g., ensuring the assessment is not measuring a different construct). Further, the lack of additional measures of algebraic knowledge prevents us from providing a benchmark for assessing the strength of the correlation of equivalence knowledge with algebraic understanding. In future work, a comprehensive pretest of algebra knowledge would go much further both in terms of confirming the differences in skills among the two cohorts and in terms of charting the correlation between level of algebra proficiency and equivalence knowledge. These issues somewhat limit the conclusions we can draw, particularly in terms of the assessment's utility in correlational data analysis. However, to our knowledge, there is currently no other existing psychometrically validated criterion measure for assessing knowledge of math equivalence. Our work is intended to push the field on this end, and future work is needed to corroborate our inferences.

Second, based on the previous measurement studies on this assessment (Matthews et al., 2012; Rittle-Johnson et al., 2011), we opted to use a construct-modeling approach with a one-parameter Rasch model. These methodological decisions were justified given our research aims, but we acknowledge that there are additional or alternative techniques that may enhance the measurement development process. For example, it is certainly possible that a two-parameter model would result in better model fit had we gathered a significantly larger sample that would allow us to use such a model. In the future, using one form of the assessment rather than two separate forms would reduce the sample size necessary to use a two-parameter model. Moreover, our analysis also does not allow us to directly compare the item difficulty estimates in this sample to those found in previous elementary-school samples. Thus, although we can examine whether performance in this older group supports the construct map and hypothesized order of difficulties, we cannot make explicit claims regarding the similarity of measurement properties across younger and older samples from different studies. Studies designed explicitly to facilitate vertical scaling across age groups would add more clarity on this end. Such studies could be specifically designed to deal with the fact that items that provide little information for one cohort, because they are very difficult or very easy for that cohort, might provide considerable information for another.

Third, the generalizability of the results remains unknown given several design decisions. We used a convenience sample of pre-algebra and algebra students from two schools within the same geographical region. We did not collect a large array of individual-level demographic characteristics, and this limited our understanding of our sample's representativeness to the larger population. Future work with more diverse populations who are served with diverse curricula is necessary to get a measure of the generalizability of our finding. Also, we used a cross-sectional design allowing us to note differences between pre-algebra and algebra students. However, longitudinal studies will be necessary to track changes in math equivalence understanding over time. Moreover, we administered the assessment in one shot as part of a measurement research study. Future research should investigate its potential use for formative assessment in real classrooms to identify students with weak understanding and to assess changes in knowledge in response to intervention. Finally, we showed that math equivalence knowledge is related to students' conceptual interpretations of an algebraic expression, but we relied on a single item to assess these interpretations. Future research could examine whether knowledge of math equivalence predicts performance on a more comprehensive assessment of variable understanding.

Given the push to make algebra accessible to all students, it is imperative to measure emerging algebraic knowledge with valid, comprehensive assessments. In the current measurement endeavor, we did just that—we focused on the assessment of math equivalence knowledge beyond elementary school and provided empirical support for the link between knowledge of equivalence and formal algebraic reasoning in middle school students.

References

Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school through college.* Washington, DC: U.S. Department of Education.

Alibali, M. W. (1999). How children change their minds: Strategy change can be gradual or abrupt. *Developmental Psychology, 35,* 127–145. http://dx.doi.org/10.1037/0012-1649.35.1.127

Alibali, M. W., Knuth, E. J., Hattikudur, S., McNeil, N. M., & Stephens, A. C. (2007). A longitudinal examination of middle school students' understanding of the equal sign and equivalent equations. *Mathematical Thinking and Learning, 9,* 221–247. http://dx.doi.org/10.1080/10986060701360902

Baroody, A. J., & Ginsburg, H. P. (1983). The effects of instruction on children's understanding of the "equals" sign. *The Elementary School Journal, 84,* 199–212. http://dx.doi.org/10.1086/461356

Behr, M., Erlwanger, S., & Nichols, E. (1980). How children view the equal sign. *Mathematics Teacher, 92,* 13–15.

Blanton, M., Stephens, A., Knuth, E., Gardiner, A. M., Lsler, I., & Kim, J. S. (2015). The development of children's early algebraic thinking: The impact of a comprehensive early algebra intervention in third grade. *Journal for Research in Mathematics Educaticn, 46,* 39–87. http://dx.doi.org/10.5951/jresematheduc.46.1.0039

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum.

Booth, J. L., & Davenport, J. D. (2013). The role of problem representation and feature knowledge in algebraic equation-solving. *The Journal of Mathematical Behavior, 32,* 415–423. http://dx.doi.org/10.1016/j.jmathb.2013.04.003

Carpenter, T. P., Franke, M. L., & Levi, L. (2003). *Thinking mathematically: Integrating arithmetic and algebra in elementary school.* Portsmouth, NH: Heinemann.

Charles, R. I. (2005). Big ideas and understandings as the foundation for elementary and middle school mathematics. *National Council for Supervisors of Mathematics: Journal of Mathematics Education Leadership, 8,* 9–24.

Chesney, D. L., McNeil, N. M., Brockmole, J. R., & Kelley, K. (2013). An eye for relations: Eye-tracking indicates long-term negative effects of operational thinking on understanding of math equivalence. *Memory & Cognition, 41,* 1079–1095. http://dx.doi.org/10.3758/s13421-013-0315-8

Falkner, K. P., Levi, L., & Carpenter, T. P. (1999). Children's understanding of equality: A foundation for algebra. *Teaching Children Mathematics, 6,* 232–236.

Fyfe, E. R., DeCaro, M. S., & Rittle-Johnson, B. (2014). An alternative time for telling: When conceptual instruction prior to problem solving improves mathematical knowledge. *The British Journal of Educational Psychology, 84,* 502–519. http://dx.doi.org/10.1111/bjep.12035

Fyfe, E. R., McNeil, N. M., & Borjas, S. (2015). Benefits of "concreteness fading" for children's mathematics understanding. *Learning and Instruction, 35,* 104–120. http://dx.doi.org/10.1016/j.learninstruc.2014.10.004

Fyfe, E. R., & Rittle-Johnson, B. (2016). Feedback both helps and hinders learning: The causal role of prior knowledge. *Journal of Educational Psychology, 108,* 82–97. http://dx.doi.org/10.1037/edu0000053

Fyfe, E. R., Rittle-Johnson, B., & DeCaro, M. S. (2012). The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters. *Journal of Educational Psychology, 104,* 1094–1108. http://dx.doi.org/10.1037/a0028389

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139–164. http://dx.doi.org/10.1177/014662168500900204

Herscovics, N., & Kieran, C. (1980). Constructing meaning for the concept of equation. *Mathematics Teacher, 73,* 572–580.

Herscovics, N., & Linchevski, L. (1994). A cognitive gap between arithmetic and algebra. *Educational Studies in Mathematics, 27,* 59–78. http://dx.doi.org/10.1007/BF01284528

Jacobs, V. R., Franke, M. L., Carpenter, T. P., Levi, L., & Battey, D. (2007). Professional development focused on children's algebraic reasoning in elementary school. *Journal for Research in Mathematics Education, 38,* 258–288.

Jones, I., Inglis, M., Gilmore, C., & Dowens, M. (2012). Substitution and sameness: Two components of a relational conception of the equals sign. *Journal of Experimental Child Psychology, 113,* 166–176. http://dx.doi.org/10.1016/j.jecp.2012.05.003

Kieran, C. (1981). Concepts associated with the equality symbol. *Educational Studies in Mathematics, 12,* 317–326. http://dx.doi.org/10.1007/BF00311062

Kinzel, M. T. (1999). Understanding algebraic notation from the students' perspective. *Mathematics Teacher, 92,* 436–442.

Knuth, E. J., Alibali, M. W., McNeil, N. M., Weinberg, A., & Stephens, A. C. (2005). Middle school students' understanding of core algebraic concepts: Equality and variable. *International Reviews on Mathematical Education, 37,* 1–9.

Knuth, E. J., Stephens, A. C., McNeil, N. M., & Alibali, M. W. (2006). Does understanding the equal sign matter? Evidence from solving equations. *Journal for Research in Mathematics Education, 37,* 297–312.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4757-4310-4

Küchemann, D. (1978). Children's understanding of numerical variables. *Mathematics in School, 7,* 23–26.

Li, X., Ding, M., Capraro, M. M., & Capraro, R. M. (2008). Sources of differences in children's understandings of mathematical equality: Comparative analysis of teacher guides and student texts in China and the United States. *Cognition and Instruction, 26,* 195–217. http://dx.doi.org/10.1080/07370000801980845

Linacre, J. M. (2013). *Winsteps Version 3.80.1 Computer Software*. Beaverton, OR. Retrieved from http://www.winsteps.com

Linacre, J. M. (2016). *Winsteps Rasch measurement computer program user's guide*. Beaverton, OR: Winsteps. http://www.winsteps.com/winman/index.htm?copyright.htm

Linchevski, L., & Herscovics, N. (1996). Crossing the cognitive gap between arithmetic and algebra: Operating on the unknown in the context of equations. *Educational Studies in Mathematics, 30,* 39–65. http://dx.doi.org/10.1007/BF00163752

Lindvall, C. M., & Ibarra, C. G. (1980). Incorrect procedures used by primary grade pupils in solving open addition and subtraction sentences. *Journal for Research in Mathematics Education, 11,* 50–62. http://dx.doi.org/10.2307/748732

Lucariello, J., Tine, M. T., & Ganley, C. M. (2014). A formative assessment of students' algebraic variable misconceptions. *The Journal of Mathematical Behavior, 33,* 30–41. http://dx.doi.org/10.1016/j.jmathb.2013.09.001

MacGregor, M., & Stacey, K. (1997). Students' understanding of algebraic notation. *Educational Studies in Mathematics, 33,* 1–19. http://dx.doi.org/10.1023/A:1002970913563

Matthews, P., & Rittle-Johnson, B. (2009). In pursuit of knowledge: Comparing self-explanations, concepts, and procedures as pedagogical tools. *Journal of Experimental Child Psychology, 104,* 1–21. http://dx.doi.org/10.1016/j.jecp.2008.08.004

Matthews, P., Rittle-Johnson, B., McEldoon, K., & Taylor, R. (2012). Measure for measure: What combining diverse measures reveals about children's understanding of the equal sign as an indicator of mathematical equality. *Journal for Research in Mathematics Education, 43,* 316–350. http://dx.doi.org/10.5951/jresematheduc.43.3.0316

McNeil, N. M., & Alibali, M. W. (2005). Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. *Child Development, 76,* 883–899. http://dx.doi.org/10.1111/j.1467-8624.2005.00884.x

McNeil, N. M., Fyfe, E. R., & Dunwiddie, A. E. (2015). Arithmetic practice can be modified to promote understanding of mathematical equivalence. *Journal of Educational Psychology, 107,* 423–436. http://dx.doi.org/10.1037/a0037687

McNeil, N. M., Fyfe, E. R., Petersen, L. A., Dunwiddie, A. E., & Brletic-Shipley, H. (2011). Benefits of practicing 4 = 2 + 2: Nontraditional problem formats facilitate children's understanding of mathematical equivalence. *Child Development, 82,* 1620–1633. http://dx.doi.org/10.1111/j.1467-8624.2011.01622.x

McNeil, N. M., Grandau, L., Knuth, E. J., Alibali, M. W., Stephens, A. C., Hattikudur, S., & Krill, D. E. (2006). Middle-school students' understanding of the equal sign: The books they read can't help. *Cognition and Instruction, 24,* 367–385. http://dx.doi.org/10.1207/s1532690xci2403_3

McNeil, N. M., Rittle-Johnson, B., Hattikudur, S., & Petersen, L. A. (2010). Continuity in representation between children and adults: Arithmetic knowledge hinders undergraduates' algebraic problem solving.

Journal of Cognition and Development, 11, 437–457. http://dx.doi.org/10.1080/15248372.2010.516421

McNeil, N. M., Weinberg, A., Hattikudur, S., Stephens, A. C., Asquith, P., Knuth, E. J., & Alibali, M. W. (2010). A is for apple: Mnemonic symbols hinder the interpretation of algebraic expressions. *Journal of Educational Psychology, 102,* 625–634. http://dx.doi.org/10.1037/a0019105

Mix, K. S. (1999). Preschoolers' recognition of numerical equivalence: Sequential sets. *Journal of Experimental Child Psychology, 74,* 309–332. http://dx.doi.org/10.1006/jecp.1999.2533

Molina, M., & Ambrose, R. (2008). From an operational to a relational conception of the equal sign: Third graders' developing algebraic thinking. *Focus on Learning Problems in Mathematics, 30,* 61–80.

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus User's Guide* (7th ed.). Los Angeles, CA: Author.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics.* Reston, VA: NCTM.

National Governors Association Center for Best Practices and the Council of Chief State School Officers. (2010). *Common Core state standards for mathematics.* Washington, DC: NGA Center & CCSSO. Retrieved from http://www.corestandards.org/wp-content/uploads/Math_Standards1.pdf

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* Washington, DC: U.S. Department of Education.

Perry, M. (1991). Learning and transfer: Instructional conditions and conceptual change. *Cognitive Development, 6,* 449–468. http://dx.doi.org/10.1016/0885-2014(91)90049-J

Powell, S. R., & Fuchs, L. S. (2010). Contribution of equal-sign instruction beyond word-problem tutoring for third-grade students with mathematics difficulty. *Journal of Educational Psychology, 102,* 381–394. http://dx.doi.org/10.1037/a0018447

Powell, S. R., Kearns, D. M., & Driver, M. K. (2016). Exploring the connection between arithmetic and pre-algebraic reasoning at first and second grade. *Journal of Educational Psychology, 108,* 943–959. http://dx.doi.org/10.1037/edu0000112

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests.* Chicago, IL: MESA Press. (Original work published 1960)

Renwick, E. (1932). Children's misconceptions concerning the symbols for mathematical equality. *The British Journal of Educational Psychology, 2,* 173–183. http://dx.doi.org/10.1111/j.2044-8279.1932.tb02743.x

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development, 77,* 1–15. http://dx.doi.org/10.1111/j.1467-8624.2006.00852.x

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology, 91,* 175–189. http://dx.doi.org/10.1037/0022-0663.91.1.175

Rittle-Johnson, B., Matthews, P. G., Taylor, R. S., & McEldoon, K. (2011). Assessing knowledge of mathematical equivalence: A construct modeling approach. *Journal of Educational Psychology, 103,* 85–104. http://dx.doi.org/10.1037/a0021334

Rosnick, P. (1981). Some misconceptions concerning the concept of variable. *Mathematics Teacher, 74,* 418–420.

Seo, K. H., & Ginsburg, H. P. (2003). "You've got to carefully read the math sentence . . .": Classroom context and children's interpretations of the equal sign. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills: Constructing adaptive expertise.* Mahwah, NJ: Lawrence Erlbaum.

Sfard, A., & Linchevski, L. (1994). The gains and pitfalls of reification: The case of algebra. *Educational Studies in Mathematics, 26,* 191–228. http://dx.doi.org/10.1007/BF01273663

Sherman, J., & Bisanz, J. (2009). Equivalence in symbolic and nonsymbolic contexts: Benefits of solving problems with manipulatives. *Journal of Educational Psychology, 101,* 88–100. http://dx.doi.org/10.1037/a0013156

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling, 3,* 25–40. http://dx.doi.org/10.1080/10705519609540027

Steinberg, R. M., Sleeman, D. H., & Ktorza, D. (1991). Algebra students' knowledge of equivalence of equations. *Journal for Research in Mathematics Education, 22,* 112–121. http://dx.doi.org/10.2307/749588

Stephens, A., Blanton, M., Knuth, E., Isler, I., & Gardiner, A. M. (2015). Just say yes to early algebra! *Teaching children mathematics, 22,* 92–101. http://dx.doi.org/10.5951/teacchilmath.22.2.0092

Van Amerom, B. A. (2003). Focusing on informal strategies when linking arithmetic to early algebra. *Educational Studies in Mathematics, 54,* 63–75. http://dx.doi.org/10.1023/B:EDUC.0000005237.72281.bf

Weaver, J. F. (1973). The symmetric property of the equality relation and young children's ability to solve open addition and subtraction sentences. *Journal for Research in Mathematics Education, 4,* 45–56. http://dx.doi.org/10.2307/749023

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97–116. http://dx.doi.org/10.1111/j.1745-3984.1977.tb00031.x

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8,* 370. http://www.rasch.org/rmt/rmt83b.htm

Appendix

Math Equivalence Assessment Items

Number	Level	Form 2 item	Form 1 item
		Equation-structure items (ST)	
1[a]	1	8 = 5 + 10 (true or false?)	8 = 5 + 10 (true or false?)
2	2	3 = 3 (true or false?)	8 = 8 (true or false?)
3	2	6 = 6 + 0 (true or false?)	4 = 4 + 0 (true or false?)
4	2	7 = 3 + 4 (explain why true)	8 = 5 + 3 (explain why true)
5[a]	3	31 + 16 = 16 + 31 (true or false?)	31 + 16 = 16 + 31 (true or false?)
6	3	7 + 6 = 6 + 6 + 1 (true or false?)	3 + 1 = 1 + 1 + 2 (true or false?)
7[a]	3	6 + 4 = 5 + 5 (explain why true)	6 + 4 = 5 + 5 (explain why true)
8	4	67 + 86 = 68 + 85 (explain why true)	89 + 44 = 87 + 46 (explain why true)
9	4	8 + 2 + __ = 10 + __	1 + 9 + __ = 10 + __
10	4	If 17 + 12 = 29, does 17 + 12 + 8 = 29 + 8? (explain why)	If 25 + 14 = 39, does 25 + 14 + 7 = 39 + 7? (explain why)
11	4	If 2 × 3 = 6, does 2 × 3 × 4 = 6 × 4?	If 3 × 5 = 15, does 3 × 5 × 6 = 15 × 6?
12[a]	4	If 76 + 49 = 121, does 76 + 49 − 9 = 121 − 9? (explain without subtracting)	If 76 + 49 = 121, does 76 + 49 − 9 = 121 − 9? (explain without subtracting)
13	4	2 × __ = 58, 8 × 2 × __ = 8 × 58 (why same number goes in the blanks)	3 × __ = 45, 7 × 3 × __ = 7 × 45 (why same number goes in the blanks)
		Equal-sign items (ES)	
14	1	6 + 4 (identify an equal pair)	3 + 6 (identify an equal pair)
15	2	5 cents __ one nickel (select symbol that shows they are the same)	10 cents __ one dime (select symbol that shows they are the same)
16	3	Is "the same as" a good definition of the equal sign?	Is "two amounts are the same" a good definition of the equal sign?
17[a]	4	Which is best definition of equal sign?	Which is best definition of equal sign?
18[a]	4	1 quarter = 25 pennies (define)	1 quarter = 25 pennies (define)
19[a]	4	What does the equal sign mean?	What does the equal sign mean?
		Open equation-solving items (OE)	
20	1	4 + __ = 8	__ + 5 = 9
21	2	8 = 6 + __	7 = __ + 3
22[a]	3	3 + 6 = 8 + __	3 + 6 = 8 + __
23	3	__ + 2 = 6 + 4	5 + __ = 6 + 2
24[a]	3	7 + 6 + 4 = 7 + __	7 + 6 + 4 = 7 + __
25[b]	3	8 + __ = 8 + 6 + 4	__ = 8 + 5 + 9
26	3	6 − 4 + 3 = __ + 3	8 + 5 − 3 = 8 + __
27[a]	4	67 + 84 = __ + 83	67 + 84 = __ + 83
28	4	43 + __ = 48 + 76	__ + 55 = 37 + 54
29	3	10 = z + 6	13 = n + 5
30[a]	4	$c + c + 4 = 16$	$c + c + 4 = 16$
31	4	$m + m + m = m + 12$	$z + z + z = z + 8$

[a] Indicates anchor items. [b] Indicates Form 1 item was intended to be __ + 9 = 8 + 5 + 9, but misprint renders it Level 2. All "Levels" were assigned a priori and hypothesized based on the construct map.

Received July 11, 2016
Revision received March 30, 2017
Accepted April 7, 2017 ■

Journal of Educational Psychology
2018, Vol. 110, No. 1, 102–118

0022-0663/18/$12.00 http://dx.doi.org/10.1037/edu0000201

A Double Dose of Disadvantage: Language Experiences for Low-Income Children in Home and School

Susan B. Neuman
New York University

Tanya Kaefer
Lakehead University

Ashley M. Pinkham
West Texas A&M University

There is a virtual consensus regarding the types of language processes, interactions, and material supports that are central for young children to become proficient readers and writers (Shanahan et al., 2008). In this study, we examine these supports in both home and school contexts during children's critical transitional kindergarten year. Participants were 70 children living in 2 different communities: neighborhoods of concentrated poverty (i.e., poverty rates over 40%) and borderline neighborhoods (i.e., poverty rates of 20–40%). From an ecological perspective, our goal was to examine the quantity and quality of knowledge-building supports in these contexts, and their relationship to children's school readiness outcomes. Interactive parent-child tasks were designed to elicit child-directed language in the home, while naturalistic observations in the kindergarten classrooms captured teachers' child-directed language. Children living in concentrated poverty were more likely to experience language of more limited complexity and diversity in both home and kindergarten contexts as compared to children living in borderline communities. We argue that the "double dose of disadvantage" in the language supports children receive at home and at school may affect their school readiness in significant, yet distinct, ways.

Educational Impact and Implications Statement

Children's early exposure to a rich set of language practices is critical for their later reading success. Nevertheless, this study shows that children from poor neighborhoods are likely to receive less complex language and fewer knowledge-building opportunities from adults in home and school, constituting a "double dose of disadvantage" in their kindergarten year. These results suggest that children will need a more expansive approach to intervention, involving both families and teachers in language-building instruction to overcome these early disadvantages to better ensure their opportunity to learn.

Keywords: language, literacy, poverty, school readiness, kindergarten

Accumulating evidence indicates that children's academic achievement is predicted not only by their family's socioeconomic status, but additionally and powerfully by the average socioeconomic status (SES) of their school (Buckingham, Wheldall, & Beaman-Wheldall, 2013). Despite expanding options for urban schooling, more than 70% of children in public schools still attend one zoned for their neighborhood (Bischoff & Reardon, 2014). This decision has serious consequences for children's educational opportunities. For example, in one meta-analysis involving over 100,000 children and 74 independent samples, Sirin (2005) found that a school's SES (i.e., the proportion of students who are eligible for free and reduced lunch) had a larger and potentially more deleterious effect on achievement than student-level SES. Although methodological characteristics, such as the type of SES measure moderated the magnitude of the relationship, the overall effect size reflected a medium level of association between SES and academic achievement at the student level and a large degree of association at the school level.

As many studies attest, however, neither student-level nor school socioeconomic status typically serve as direct, causal factors in children's academic achievement (e.g., Duncan & Murnane, 2011; Mayer, 1998). Instead, SES may affect children's day-to-day quality of life, including their access to resources associated with learning, the quality of local services, and their socialization by both familiar adults and larger social networks (Neuman & Celano,

This article was published Online First April 13, 2017.

Susan B. Neuman, Steinhardt School of Culture, Education, and Human Development, New York University; Tanya Kaefer, Faculty of Education, Lakehead University; Ashley M. Pinkham, Department of Psychology, Sociology and Social Work, West Texas A&M University.

Correspondence concerning this article should be addressed to Susan B. Neuman, Steinhardt School of Culture, Education, and Human Development, New York University, 239 Greene Street, Room 626, New York, NY 10003. E-mail: sbneuman@nyu.edu

2012). Limited access to these resources may, in turn, contribute to SES-related differences in children's achievement.

SES may also indirectly influence achievement by impacting children's language skills as they reach school age. Although linguists have shown definitively that all biologically intact children have well-developed language (Brown, 1973; Chomsky, 1968) and "ways with words" (Heath, 1983), at the same time studies have reported striking differences across SES groups on a variety of measures including the mastery of complex sentence structures (Vasilyeva, Waterfall, & Huttenlocher, 2008). Converging evidence from both experimental and longitudinal studies demonstrates stark differentials in phonological awareness, vocabulary, and oral language comprehension between low- and middle-income children prior to entering school (Cunningham & Stanovich, 1997; Hart & Risley, 1995). In fact, Fernald, Marchman, and Weisleder (2013) recently reported significant SES-related disparities in vocabulary, and language processing efficiency in infants as early as 18 months of age. By 24 months of age, children from higher-SES backgrounds demonstrated a 6-month gap advantage compared to children from lower-SES backgrounds. Early differences in foundational language and literacy skills, including phonological sensitivity to sounds and the size and depth of vocabulary knowledge, may serve as a direct precursor of later SES-related achievement gaps. Moreover, such differences may support what Mol and Bus (2011) called a "spiral of causality": children who are more proficient in school-related language skills become increasingly capable in academic reading, spelling and comprehension, whereas those who are less proficient may spiral downward over time.

The Transition From Home to School

Although student-level SES factors (e.g., family income, parent education, occupation) are necessary for our understanding of children's early literacy development, they may be insufficient for providing a complete picture of how children become (or do not become) proficient and successful readers. Because inequality is often organized or clustered in social settings like neighborhoods and schools (Lareau & Goyette, 2014), where a child lives may add an additional dimension of stratification. In a recent analysis using the *Panel Study of Income Dynamics*, Sharkey (2013) showed remarkable stability of neighborhood disadvantage as well as neighborhood advantage across multiple generations. Such findings indicate that progress toward racial and economic equality may be "stuck in place."

This persistent inequality may not bode well for low-income children's achievement. A large body of research (e.g., Brooks-Gunn, Duncan, & Aber, 1997; Massey, 2007) suggests that children in higher poverty neighborhoods are more likely to attend troubled schools, have less qualified teachers, and have lower levels of academic achievement as compared to their peers in more affluent neighborhoods. In this respect, children may face a "double dose of disadvantage" as they begin to participate in school settings. More specifically, children may go from one context (the home) with limited physical and psychological resources for this type of language and early print experiences to another context (the school environment) with similar constraints. In this manner, children from lower-SES backgrounds may spend the majority of their waking hours experiencing fewer opportunities for language and literacy development. Although these children may have unique linguistic strengths that serve them well in their immediate settings, the double dose of disadvantage may result in relatively weak academic language skills that serve as obstacles to overall school achievement (Hoff, 2013).

In the present paper, we address this possible "double dose of disadvantage" by following a cluster sample of 70 children who have recently made the transition from preschool to kindergarten. Thirty-five of the families and their children came from neighborhoods considered to reflect concentrated poverty (i.e., 40% or more of residents identified as poor), while the remaining half came from contiguous "borderline" communities that are more demographically diverse and largely working-class. Children from the concentrated poverty neighborhoods had all attended Head Start in the previous year. Consistent with the *Head Start Impact Study* (Puma, Bell, Cook, & Heid, 2010), these children demonstrated substantial progress in school readiness skills, including language, vocabulary, and letter-knowledge abilities.

In this study, we followed these children through their kindergarten year, conducting targeted observations in both home and school settings. Our goal was to address the extent to which these children continued to receive language and print supports in the year after receiving early intervention. At the same time, we conducted similar observations for the 35 children living in the borderline working-class communities. By comparing the supports children received across communities and contexts, we attempted to understand (a) how adult child-directed language from home and school may contribute to children's learning trajectories, and (b) how home and school SES factors in neighborhoods may be inextricably connected.

Physical and Psychological Supports for Early Learning

Underlying the concept of the environmental support hypothesis is the theoretical premise that environment plays a critical role in children's development. Specifically, Bronfenbrenner and Morris's bioecological model posits that human development results from a complex interplay of Process × Person × Context × Time. Proximal processes, such as interactions with adults and access to learning support materials, serve as the primary means with which young children learn in home and classroom environments (see Bronfenbrenner & Morris, 1998, for further review).

There is virtual consensus regarding the types of proximal processes that are central for children to become proficient readers and writers (see Shanahan et al., 2008). In particular, children need a rich conceptual knowledge base (Harris, Golinkoff, & Hirsh-Pasek, 2011), a broad and deep vocabulary (Beck & McKeown, 2007), and strong verbal reasoning skills (Brown, Roediger, & McDaniel, 2014) in order to understand the messages that are conveyed through print. In addition to serving as the foundation for later print experiences, these language skills also support the development of code-related skills, including phonological awareness, the alphabetic principle, and the many systematic correspondences between sounds and spellings (Shanahan et al., 2008).

But to attain these skills, children require many opportunities with supportive adults who act as social and psychological resources that provide information through extensive feedback and demonstrations (Neuman & Carta, 2011). Interacting with adults who produce more complex utterances and use a greater variety of syntactic structures in their spontaneous speech may especially promote children's expressive language and syntactic development (Hoff, Rumiche, Burridge, Ribot, & Welsh, 2014). Caregivers' diversity of language input, or their use of a variety words, may also have a facilitative effect on children's language and literacy development. Weizman and Snow (2001), for example, found that mothers' greater use of sophisticated words during everyday activities later predicted their children's language learning (see also Rowe, Raudenbush, & Goldin-Meadow, 2012). Yet it is not an entirely one-sided affair: responding contingently to what children are saying or asking, as well as extending what children are curious about learning, also fosters language development and learning (Landry, Smith, Swank, Assel, & Vellet, 2001; Mol & Neuman, 2014).

Although the lexical features (e.g., quantity, complexity) of caregiver child-directed speech clearly influence children's language development, the *content* of such speech is also highly important. Language is a key mechanism for conveying important conceptual understandings and knowledge. For example, children may learn broad generalizations about the taxonomic categories in their world through conversations with adults (see Gelman & Kalish, 2006). Generic noun phrases, such as "cats have whiskers," describe shared properties of a category (e.g., cats as a category, whiskers as a property of the category) and express which properties may be extended to new category exemplars (Brandone, Leslie, Cimpian, & Gelman, 2012). Such generic noun phrases appear frequently in naturally occurring conversations between supportive adults and young children and are thought to play an instrumental role in children's conceptual development and reasoning. In fact, children as young as 2 [1/2] years of age will produce such generic noun phrases in a context that encourages such focus (Gelman & Raman, 2003).

Furthermore, statements that engage children in representational thinking can be a source of abstract thinking (LaRusso et al., 2016). Such conversations include anticipating events, reconstructing past events, and transcending from the immediate present to the symbolic level. For example, asking children to describe an experience or a story requires that they separate from the immediate experience in order to think about the previous event, thereby placing cognitive demands to transcend the present to deal with the past (van Tuijl & Leseman, 2004). Sigel (1982) described the phenomenon as "distancing" and demonstrated that the accumulation of these experiences is positively associated with children's ability to classify and represent knowledge using oral and written symbol systems.

In sum, children's early exposure to a rich set of language practices may set in motion the processes that they will need for learning to read, including the vocabulary, and background knowledge necessary for language and reading comprehension (Neuman & Celano, 2012). Consequently, children who have limited experience with these kinds of linguistic interactions may have fewer opportunities to engage in higher-order exchanges of cognitive challenge valued in school and are likely to be at a disadvantage later on (Bailey, Duncan, Odgers, & Yu, 2015). Crucially, such differences in the quantity and quality of adults' child-directed language are strongly associated with children's socioeconomic background. They may also partially account for later SES-related disparities in children`s lexical and grammatical development both within and between SES groups (Hart & Risley, 1995; Hoff, 2006). These SES-related differences not only influence children's immediate skill-building, but their ability to process and learn from future language input as well. If these children also experience school contexts in which teachers provide less language-advancing input and fewer opportunities for knowledge-building supports, they may be less likely to progress over the course of the school year as compared to children with teachers who use academically productive talk more strategically and deliberately for learning and reading development (Stanovich, 1986). Taken together, this may constitute a double dose of disadvantage for low-income children.

In this study, we examine these language-advancing supports in both home and school for children in the critical transitional year from preschool to kindergarten. Given the important role that adults' child-directed language plays in fostering children's language skills, vocabulary, and background knowledge, we asked the following questions:

(a) To what extent are there differences in the quantity and quality of parent-child interactions for children living in concentrated poverty as compared to their working-class peers?

(b) When children living in concentrated poverty and borderline communities transition to kindergarten, are there differences in the characteristics of teachers' use of child-directed language supports?

(c) To what extent do differences in language and knowledge-building supports at home and in the classroom predict children's early academic achievement?

Method

Data Sources

This study is a follow-up of a large-scale, cluster randomized controlled trial ($n = 585$) measuring the effects of a supplemental curriculum, the *World of Words*, on children's vocabulary and early literacy development (Neuman, Newman, & Dwyer, 2011). The original study took place in 25 classrooms throughout a county-wide Head Start program in Detroit, a severely economically depressed Midwestern city, and its surrounding area. At the time of the study, the city experienced 25% unemployment and an overall child poverty rate of 67%. Parent consent for the program, obtained through the Head Start County Office was 100%.

In addition to the typical business-as-usual control groups of Head Start classrooms, we also wanted to examine how students' growth in vocabulary compared to those who came from neighborhoods with less severe poverty. To do so, we included an additional comparison group of children ($n = 221$) from three smaller contiguous areas described by Jargowsky (2014) as borderline neighborhoods, in which the poverty rate generally ranges from 20–40%. Although still showing signs of distress, such neighborhoods tend to be more working-class or lower-middle-income communities. Children in this group were not eligible for Head Start (i.e., income was above the eligibility criterion) and instead attended a state-funded prekindergarten program ($n = 24$). Eligibility for the state-funded program was based on evidence of risk factors, including income (i.e., low-income but above federal income poverty guidelines) and 22 other possible risk factors (e.g., maternal education, single parent household, speech delays, ELL status; Michigan Great Start Readiness Program, 2012). Children

Table 1
Sample Characteristics of the Original 2011 Study

Descriptive	Head Start treatment (N = 294)	Head Start control (N = 289)	Borderline control (N = 221)
Average age in months	49	50	52
Mean WJ pretest (Sept.)	98.44	97.51	100.94
% Female*	55	51	53
% Minority*	58%	55%	39%

Note. WJ = Woodcock Johnson.
$^{*}p < .05$.

qualified for the program if they were considered at risk on the basis of at least two of these factors. Sample characteristics of the original study are shown in Table 1.

At the end of Year 1, the randomized trial showed substantial improvements by the Head Start treatment group in researcher-developed vocabulary (Cohen's d = .86) and conceptual development measures (Cohen's d = .53) as compared to the Head Start control group. In addition, treatment children showed average progress in expressive vocabulary at the conclusion of the year, as measured by the Woodcock-Johnson Picture Vocabulary test, with no significant differences between children in the Head Start treatment group (100.15, SD = 13.81) and those in the state-funded program (101.87, SD = 11.82; see Neuman et al., 2011).

Participants

In the present analysis, our goal was to examine the continuous progress for a subset of children from the concentrated poverty and borderline communities. Two participants (i.e., one boy and one girl) were randomly selected from each of the Head Start classrooms (n = 50) and state-funded preschool classrooms (n = 48). Of the 98 children identified, contact information was available for 80 of the families; this was equally distributed across the two groups. Following an initial phone conversation with families, five children were eliminated either due to limited English proficiency (two children) or expected moves out of the state (three children).

Letters were then sent to the families of the remaining 75 children describing the study and our interest in better understanding children's transition to kindergarten. In the informed consent, we outlined our requests for a series of four home visits throughout the year and requested permission to observe children's everyday activities in their kindergarten classrooms on approximately four occasions. Seventy families returned informed consent and agreed to participate (i.e., 93% response rate). Participating families were given a $25 gift card and a new book for the child's personal library after each home visit; teachers received a small library for their classroom at the end of the study.

Of the 70 children in our final sample, 35 had attended Head Start, and lived in the economically depressed urban community, and were identified as low-income (N = 18, treatment; N = 17, control). The remaining 35 children had attended the state-funded pre-K program, and lived in one of the three borderline neighborhoods, and were identified as working-class. For both groups, participants' scores on the expressive vocabulary measure did not significantly differ from their larger group (i.e., Head Start treatment, Head Start control, or baseline comparison) at the end of their pre-K year.

Demographic information of the families and the neighborhoods is provided in Tables 2 and 3. As shown, there were significant differences between groups on parental age, education, and income. The difference in the composition of the neighborhoods, such as average income, density of poverty was substantial as well. Low-income parents were younger, poorer, and more likely to be of minority status than working-class parents. Home environ-

Table 2
Demographic Characteristics of Families and Neighborhoods

Variable	Concentrated poverty (n = 35)	Borderline communities (n = 35)
Child's age (in months)	61.68 (.37)	59.28 (.65)
Child's gender		
Male	46%	40%
Female	54%	60%
Parent age (in years)		
Mother*	33.94 (8.86)	38.59 (6.22)
Father*	36.33 (9.24)	41.81 (10.57)
Parent education		
Mother**		
Did not complete high school	15%	0%
High school diploma	59%	16%
Some college	26%	28%
Bachelor's degree	0%	15%
Post-graduate degree	0%	41%
Father**		
Did not complete high school	7%	0%
High school diploma	63%	15%
Some college	30%	30%
Bachelor's degree	0%	15%
Post-graduate degree	0%	41%
Family income**		
Less than $15,000	20%	0%
$15–39,999	51%	6%
$35–49,999	26%	18%
$50,000–74,999	3%	26%
More than $75,000	0%	50%
Family ethnicity		
Caucasian	41%	61%
African American	41%	21%
Hispanic/Latino	3%	6%
Asian	0%	6%
Bi-/multi-racial	15%	6%
H.O.M.E. (max = 55)**	42.37 (6.25)	50.63 (3.03)
School SES (% free/reduced lunch)	45%	35%
School failing to make AYP	21%	0

Note. SES = socioeconomic status; H.O.M.E. = Home Observation for the Measurement of the Environment; AYP = Adequate Year Progress.
$^{*}p < .05$. $^{**}p < .01$.

Table 3
Demographic Characteristics of Neighborhoods[a]

Neighborhood	Population	% Poverty	Average household income	Race/Ethnicity	Education (% completing at least high school degree)
Concentrated poverty					
Detroit	713,777	41.3%	$26,325	Caucasian: 12% African American: 83% Other: 5%	77.8%
Hamtramck	22,423	43.5%	$25,659	Caucasian: 50% African American: 11% Other: 39%	64.8%
Highland Park	11,776	51.1%	$18,981	Caucasian: <1% African American: 89% Other: 11%	73.8%
Borderline communities					
Taylor	63,131	21.2%	$41,933	Caucasian: 76% African American: 17% Other: 3%	82.3%
Romulus	23,989	21.4%	$44,119	Caucasian: 49% African American: 44% Other: 7%	85.7%
Sumpter	9,549	19.5%	$53,109	Caucasian: 89% African American: 11% Other: <1%	88.6%

[a] American Fact Finder (http://factfinder.census.gov).

ments, as well, were less cognitively stimulating in these poor neighborhoods. Additionally, poor children were more likely to attend schools with a higher percentage of free and reduced lunch than those children from working-class families.

Furthermore, their schools were rated more poorly than those in the working-class communities: Accountability measures indicated that over 21% of the schools in these high-poverty neighborhoods failed to meet adequate yearly progress (as identified through NCLB); in addition, composite school reports used as an additional accountability measure in the State, indicated that 45% of these schools, compared to 16% in working-class neighborhoods were in need of improvement. These demographic characteristics of the schools are consistent with previous studies (Dupere, Leventhal, Crosnoe, & Dion, 2010), which have shown how neighborhood effects are channeled through a series of related contexts (e.g., schools; classrooms) embedded in and often bounded by their neighborhood environment. In addition, they show how factors associated with children's opportunity to learn are organized or clustered in social settings like neighborhoods and schools.

Data Collection

The goal of our project was to better understand the home and school transition for these children coming from two different environmental settings: (a) neighborhoods in which children grew up in concentrated poverty, characterized by the poverty level in the community exceeding a threshold level of 40% or more; and (b) borderline communities that were affected by the Great Recession and had poverty levels ranging from 20% to 40%, but continued to have greater job stability than neighborhoods characterized by concentrated poverty. The purpose of our data collection, therefore, was to examine the types of language and literacy processes typically valued in school in both contexts.

To do so, we scheduled four hour-long visits in family homes, starting in the fall and spaced throughout the school year. During each visit, we administered both standardized assessments and tasks specifically designed to engage parents and children in joint interactions (see below). Starting in late fall, we also scheduled four observations in children's kindergarten classroom; these observations were spaced approximately 6 weeks apart. Each observation consisted of a half-day classroom visit during which we recorded teachers' child-directed talk. Although the specific contexts and the occasions for recording adult/child interactions varied between home and school contexts, our goal was to examine a common set of criteria for adult child-directed talk known to promote to language and reading development. Table 4 provides an overall description of the data collected in both settings.

Home Measures

Assessments and tasks were designed to examine the degree of cognitive stimulation in the home, the quality of parent-child interactions, and children's school readiness skills. We used the following measures:

Home observation for the measurement of the environment. During the first home visit, trained research assistants administered the Home Observation for the Measurement of the Environment (H.O.M.E.; Bradley, Caldwell, & Corwyn, 2003), a widely used observational measure designed to examine the quality and quantity of cognitive stimulation and material support in children's homes. H.O.M.E. has been used extensively in research has been shown to be sensitive to increments in family income, particularly when looking at children born into poverty (Totsika & Sylva, 2004). The measure includes 55 items, clustered in eight subscales (i.e., learning materials, language stimulation, physical environment, parental responsivity, learning stimulation, modeling of social maturity, variety in experience, and acceptance of child). The

Table 4
Overview of Study Variables

Variable	Description
Home	
H.O.M.E.	An observational measure of the quality and quantity of stimulation and support in the child's home.
Parent language	
Word tokens	Total number of words spoken by parent
Lexical density	Average number of words per minute
Lexical diversity	Total number of unique word types
Sentence complexity	Average number of words per utterance
WRAT4	Reading comprehension assessment
Teacher language	
Word tokens	Total number of words spoken by teacher
Lexical density	Average number of words per minute
Lexical diversity	Total number of unique word types
Sentence complexity	Average number of words per utterance
Child outcomes	
PPVT-IV[a]	Standardized measure of receptive vocabulary
WJIII-Picture Vocabulary[b,c]	Standardized measure of expressive vocabulary
Letter/Sound identification	Researcher-developed criterion measure of letter and sound identification
WJIII-Letter/Word identification	Standardized measure of basic decoding skills
WJIII-Word attack[c]	Standardized measure of basic phonics

Note. H.O.M.E. = Home Observation for the Measurement of the Environment.
[a] WRAT4 = Wide Range Achievement Test. [b] PPVT-IV = Peabody Picture Vocabulary Test-IV. [c] WJIII = Woodcock Johnson-III.

first five observations were independently coded by two research assistants; interrater reliability was κ = .95. Once interrater reliability was established, a single coder coded subsequent observations.

Wide range achievement test. During the first visit, a second trained research assistant administered the Wide Range Achievement Test-Fourth edition (WRAT4; Wilkinson & Robertson, 2006), a brief achievement test normed to measure reading comprehension in teenagers and adults (up to age 94). The reported internal-rater reliability for the WRAT4 is .98. Used widely as a norm-referenced measure of reading comprehension, the WRAT has been standardized across a stratified sample including gender, race, ethnicity, and region.

The WRAT was administered to the primary caregiver (usually the mother) as an assessment of his or her reading proficiency. The assumption was that parents' own reading ability might relate to their interest and involvement in reading with their children. Furthermore, previous research has demonstrated that mothers' educational background is highly correlated with their children's school readiness skills (Denton, West, & Waltston, 2003). The WRAT took approximately 15 min to administer.

Quality of parent-child interactions. For the second and third visits, we developed two tasks designed to engage parents and children in joint activity. These activities were deliberately targeted to elicit parent child-directed language, allowing us to focus on the variety of the dyadic interactions and to examine the knowledge-building supports parents provided during these interactions. Although either parent was invited to join in the interaction, only mothers chose to participate in the activity.

The first activity, the *Literacy Prop Bag,* was administered during the second home visit. The task was initially adapted from the NICHD's Study of Early Child Care's three-bag assessment (NICHD Early Child Care Research Network, 1999) and had been used in our previous research (Mol & Neuman, 2014). The three bags in this task included: an age-appropriate narrative nonfiction book, *Have you seen bugs?* (Oppenheim, 1996); a set of realistic toy insects; and a notebook for recording children's drawing and/or writing about the activities.

Instructions were deliberately open-ended to observe the natural variation that might occur in parent-child interaction. Parents were given the first bag and asked to share the book with their children. After sharing the book, a second bag with an assortment of insect replicas related to the book was given to the dyad. Following their play, they were given a third bag containing the notebook. Dyads engaged in each activity for approximately 5 min (book: $M = 5.96$ min; toys: $M = 5.50$; notebook: $M = 3.25$). During the activity, two trained research assistants independently took informal notes on parent-child interactions and conversations. The session was also audiotaped (with parent permission).

During the third home visit, parent-child dyads completed the second activity, *Animal Bingo.* This activity involved a matching game with animal pictures. It was designed to engage parents and children in language interactions using a more structured task than the Literacy Prop Bag. Dyads were given a game-board composed of a 4 × 4 grid with illustrations of both familiar animals (e.g., pig) and less familiar animals (e.g., starfish). They were also given 16 playing cards, each one corresponding to an illustration on the game-board. Parents were asked to select playing cards one at a time and to help their children identify the specific animal without using the animals' names or providing overt cues (e.g., pointing). Children were instructed to listen carefully and try to find the animals being described. After the game was completed, the parent/child dyad was given a second game-board and a new set of 16 playing cards, and asked to play the game again. During this second, more challenging round, half of the playing cards contained illustrations of new animals, while the remaining eight cards were printed with the name of an animal. On average, each round of the collaborative game lasted 6.43 min (*SD* = 1.85). The total parent-child interaction averaged 12.90 min (*SD* = 3.97) in length. Sessions were audiotaped while research assistants took informal notes.

Child Assessments

Children's language and school readiness skills were also assessed during these home visits. Four measures were assessed once during the fall and once during the spring: receptive vocabulary knowledge, expressive vocabulary knowledge, letter-name knowledge, and letter-sound knowledge. Two additional measures were administered only in spring to assess children's initial decoding skills.

Peabody Picture Vocabulary Test-IV. Children's receptive vocabulary knowledge was measured using the Peabody Picture Vocabulary Test-IV (PPVT-IV), a vocabulary test which yields both raw scores and standard equivalent scores related to national norms (Dunn & Dunn, 2007). Form A was administered during the fall, and Form B was administered in the spring. Standard scores were used for all analyses. The reported reliability for the PPVT-IV ranges from .93 to .95.

Woodcock-Johnson III-Picture Vocabulary. Children's expressive vocabulary knowledge was measured with the Woodcock-Johnson III-Picture Vocabulary subtest (WJIII-PV; Woodcock, McGrew, & Mather, 2007). The subtest consists of 42 pictures of words that increase in difficulty. For each item, children are prompted to label the picture; administration is discontinued after failing to label six items in a row. Scaled scores based on national norms were used for all analyses. The reported reliability for the WJIII ranges from .8 to .9.

Letter names and letter sounds. We examined children's ability to name letters and identify sounds through two informal measures. Seven letters were selected based on high utility (i.e., *s, m, t, b, p, r, f*). These letters were placed on flashcards, which were presented in random order. Children were asked to identify the name of each letter as well as its sound. Children received a total correct score for both letters (maximum = 7) and sounds (maximum = 7). Cronbach's alpha indicated .95.

Woodcock-Johnson III-Letter-Word identification. This WJIII subtest assessed children's ability to name or pronounce letters of the alphabet. Both upper- and lowercase letters are used, and letters are presented in a variety of type styles. Reliability of the measure is .94. Due to the limited number of items for young children at the basal level, this measure was administered only as a spring posttest.

Woodcock-Johnson III-Word attack. This measure examines the ability to apply phonic/decoding skills to unfamiliar words (Woodcock et al., 2007). Initial items require children to produce sounds for a small set of single letters; later items require children to pronounce pseudowords of increasing complexity. Reported reliability for the measure is .90. This measure was administered only as a spring posttest.

School Measures

Kindergarten teacher questionnaire. All teachers were asked to complete a demographic survey detailing their background characteristics, teaching experience and the curricula they used in their classrooms. The survey took 10–15 min to complete.

School socioeconomic status (SES) and accountability measures. Data on percent free and reduced lunch for each school, as well as school composition, were retrieved from publicly available data through the American Factfinder website (factfinder.census.gov). Accountability information was retrieved from state report cards (www.mischooldata.org).

Classroom observations. We conducted a series of observations to examine teachers' child-directed language in the kindergarten classrooms attended by each target child. With parent permission, we contacted school administrators. With their permission, as well as that of the teacher's, we scheduled four classroom visits over the coursc of late winter and throughout the spring. In the majority of schools, there was one kindergarten teacher with one student participating in the study; however, in the remaining nine schools, there were two or more kindergarten classrooms, each with one student involved in the study. Sixty-three children (90%) attended their neighborhood elementary school. Eight of these children (11%) attended a kindergarten classroom in the same school with other targeted children. Together, kindergarten teachers from a total of 54 schools were included in the study.

Observations were conducted by 10 trained research assistants (one per classroom). Each observation began at the start of the school day and lasted for approximately three hours. During each observation, lead teachers wore a portable microphone to allow for all teacher child-directed language to be audio-recorded. At the same time, research assistants sat at the back of the room, listened to the talk, and took informal notes on a laptop computer to provide additional context for the activity or to report on any potential instructional interruptions (e.g., phone call). Together, we collected close to 12 hours of teacher child-directed talk. For additional information about classroom observations, see (Wright & Neuman, 2014).

Analyzing Adult-Child Language Interactions

All speech produced during both parent observation tasks (i.e., literacy prop bag, animal bingo) and classroom observations were transcribed in full by trained research assistants using the Codes for Human Analysis of Transcripts (CHAT) transcription system available through the Child Language Data Exchange System (CHILDES; MacWhinney, 2000). In this system, the unit of transcription is the utterance, defined as any sequence of words preceded and followed by a pause, change in conversational turn, or change in intonational pattern. To ensure consistency across transcriptions, each transcript was reviewed against the audio-recording at regular intervals. When there was any ambiguity (approximately 2% of the time), transcribers referred to contextual notes taken by the original research assistant for further clarification.

The Computerized Language Analysis (CLAN) program was then used to analyze data transcribed in the CHAT format, calculating the exact number or ratio of scores. CLAN facilitates the automatic computation of indices such as lexical density, mean length of utterances (MLU), and other measures of conversational analyses.

Coding child-directed language. Since our goal was to examine adults' use of child-directed language in home and school, we operationalized the linguistic input in these two settings in similar ways. Specifically, we focused on the quantity and the quality, or lexical richness, of the language input in these two settings.

Quantity of language input. The quantity of language was coded in two ways: word tokens and lexical density. Word tokens measured the total number of words spoken by the adult (i.e., parent or teacher) during the entire observation, independent of duration. By contrast, lexical density measured the amount of linguistic input relative to the overall duration of the interaction. Specifically, the measure calculated the average number of words produced per minute during the observed period.

Quality of language input. The quality of language input captured the extent to which adults used rich and varied language when talking to children. We operationalized quality in two ways: lexical diversity and sentence complexity. Lexical diversity addressed the overall variety of vocabulary words used in child-directed talk. Using the FREQ command in CLAN, we produced a comprehensive list of words spoken by adults, and then determined the number of different word types (i.e., different word roots) produced in each observation. All dictionary words, as well as onomatopoeic sounds were counted as words; however, unconventional or pseudowords were excluded. Morphologically inflected variants of words (e.g., read; reading) were considered a single type.

Sentence complexity was operationalized as the average number of words per adult utterance. Studies suggest that adults who speak in longer utterances may have greater opportunity to use diverse vocabulary and syntax. They may also provide children with more information in general, including information about specific word meanings (Dickinson, Hofer, Barnes, & Grifenhagen, 2014; Dickinson & Porche, 2011). The average number of words per utterance was computed using the MLU command in CLAN.

Results

These observations of home and school contexts, combined with child assessments of language and early literacy skills, allowed us to examine how language and knowledge-building supports in different environmental settings might relate to children's receptive and expressive language and early literacy skills in the important transitional year from preschool to the elementary grades. We present our results in four sections. First, we examined the differences between treatment and control within the low-income group to determine any potential carryover effects of the intervention. Next, we analyzed the features of parent child-directed language and measured whether there were differences between low-income and working-class groups. To that end, based on the high correlations among these language characteristics, we used multivariate analysis of variance (MANOVA) to examine the differences in linguistic input broadly, followed by univariate analyses to examine differences for each language characteristic. Next, we turned to the school context. We first focused on the demographic characteristics of the classrooms. Then we addressed the qualitative features of teacher child-directed language, followed by a similar set of analyses to examine differences between groups. Although students were clustered within classrooms for the initial intervention (Neuman et al., 2011), at kindergarten they were dispersed and no longer clustered within the same schools or classrooms (i.e., only six classes had two children per class). Therefore, multilevel modeling (e.g., HLM) was not considered appropriate in these analyses (Galbraith, Daniel, & Vissel, 2010; McCoach & Adelson, 2010).

Together, these analyses were designed to determine the extent to which there were differences in the quantity and quality of child-directed language from parents and teachers with children who came from low-income homes compared to their working-class peers. Finally, we conducted a series of hierarchical linear regression models to examine whether these differences in language and knowledge-building supports predicted children's early achievement.

Differences Within the Low-Income Group

Our first analysis was to examine whether there were differences within the low-income group, given that a portion of the sample had received an intervention the previous year. To conduct this analysis, we clustered the outcome variables into three categories: parent language, teacher language, and child outcomes. For each category, we conducted a MANOVA analysis, with children who had participated in the treatment or control as the independent variable. Table 5 describes these findings.

In each case, there were no significant differences between groups on parent language F (6, 24) = 1.96, p = .118; teacher language F (5, 28) = 1.64, p = .215, or child outcomes, F (6, 24) = .65, p = .691. These results indicated that the previous year's intervention did not have a carry-over effect to these outcomes.

Differences in Parent-Child Directed Language Across Communities

Table 6 describes the parent child–directed language features across the two groups as well as parents' reading comprehension scores. To conduct this analysis, we first examined whether there were differences in duration of the activity across the two groups. We found a significant difference, t (58) = 2.48, p = .016, d = .65 which suggests that parents from working-class backgrounds used more of the available time within the allowed time period of the activities than parents from the low-income backgrounds. Despite the significance of this finding, however, the overall difference was less than a minute. Next, we conducted a MANOVA to examine SES differences in parent language variables, controlling for the duration of interaction. As shown in Table 6, planned univariate analyses indicated significant effects of SES on sentence complexity (F [1, 54] = 2.27, p = .024, *partial* η^2 = .09), lexical diversity (F [1, 54] = 7.93, p = .007, *partial* η^2 = .13), and parent WRAT scores (F [1, 55] = 4.86, p = .032, *partial* η^2 = .09). Of particular note, these differences were recorded when parents in these communities engaged in *similar* activities and tasks in contrast to previous studies of language in day-to-day natural contexts. Parents in these low-income homes used shorter sentences, and fewer different words than those in the working-class community. In addition, parents who came from these low-

Table 5
Descriptive Statistics (Means and Standard Deviations) of Variables in Current Study by Participation Group in 2011 Study

Variable	Treatment (n = 18)	Control (n = 17)
Parent variables		
Duration of activity (in minutes)	15.27 (1.02)	15.28 (.37)
Word tokens (number of words)	824.82 (278.45)	806.43 (269.16)
Lexical density (words per minute)	53.15 (17.31)	51.99 (17.26)
Lexical diversity (number of word types)	207.53 (50.28)	187.50 (42.46)
Sentence complexity (words per utterance)	4.86 (.61)	4.87 (.56)
WRAT4	92.13 (11.86)	90.85 (13.67)
Teacher variables		
Duration of activity (in minutes)	104.23 (40.06)	125.97 (31.39)
Word tokens (number of words)	7978.92 (3592.59)	7570.88 (3983.86)
Lexical density (words per minute)	76.65 (23.15)	57.31 (20.27)
Lexical diversity (number of word types)	695.75 (235.36)	664.75 (183.87)
Sentence complexity (words per utterance)	5.15 (1.02)	5.24 (.90)
Outcome variables		
WJIII-Picture Vocabulary[a]	105.47 (12.43)	100.75 (6.02)
PPVT-IV[b]	101.40 (10.34)	96.44 (11.52)
Letter names	.83 (.30)	.83 (.26)
Letter sounds	.59 (.41)	.65 (.32)
WJIII-Letter/Word identification	109.13 (10.11)	112.00 (11.40)
WJIII-Word attack	81.27 (52.99)	62.38 (57.39)

Note. WRAT4 = Wide Range Achievement Test.
[a] WJIII = Woodcock Johnson-III. [b] PPVT-IV = Peabody Picture Vocabulary Test-IV.

income communities had significantly lower reading comprehension scores—in fact, approximately a full standard deviation difference—than those from the working-class communities. However, there were no significant differences in lexical density, $F(1, 54) = 2.85$, $p = .097$, or in quantity of language overall, $F(1, 54) = 2.49$, $p = .121$.

Taken together, these results show that there were significant differences in parent child-directed language and parent's own reading comprehension skills across these two communities.

Differences in Teacher-Child Directed Language Across Community

Our next step was to look at the transition to school, specifically the kindergarten classroom and the quality of teachers' child-directed language. As shown in Table 7, children from the lower-income community were likely to attend schools in which the demographic characteristics of their kindergarten teachers were not significantly different than those who attended schools in

Table 6
Descriptive Statistics (Means and Standard Deviations) of Parent and Teacher Child-Directed language

Variable	Concentrated poverty (n = 35)	Borderline communities (n = 35)
Duration of activity (in minutes)		
Parent*	15.22 (.81)	15.74 (.82)
Teacher	110.18 (35.30)	124.20 (30.40)
Word tokens (number of words)		
Parent	716.04 (293.55)	873.23 (247.82)
Teacher	7191.45 (3246.81)	8579.21 (2858.01)
Lexical density (words per minute)		
Parent	46.23 (18.07)	55.54 (15.31)
Teacher	65.12 (20.35)	70.13 (18.60)
Lexical diversity (number of word types)		
Parent**	181.93 (44.87)	223.20 (44.52)
Teacher*	650.00 (193.36)	754.63 (118.86)
Sentence complexity (words per utterance)		
Parent*	4.62 (.66)	5.11 (.65)
Teacher*	4.86 (.73)	5.52 (.89)
WRAT4		
Parent*	91.26 (11.17)	99.57 (16.46)
Teacher	—	—

Note. WRAT = Wide Range Achievement Test.
* $p < .05$. ** $p < .01$.

Table 7
Demographic Characteristics of Teachers and Schools

Variable	Concentrated poverty (n = 35)	Borderline communities (n = 35)
Number of classrooms	29	25
Full-day kindergarten	56%	63%
Part-day kindergarten	44%	37%
Teacher education		
Bachelor's degree	35%	41%
Master's degree	65%	57%
Teacher age (in years)	40.52	41.76
Years worked as a teacher		
Less than 5 years	38.5%	31%
6–10 years	15%	17%
11–15 years	11.5%	7%
16+ years	35%	45%
Years worked as kindergarten teacher		
Less than 5 years	46%	48%
6–10 years	11.5%	21%
11–15 years	23%	17%
16+ years	19%	14%
Teacher ethnicity		
Caucasian	76%	86%
African American	20%	7%
Hispanic/Latino	0%	3.5%
Asian	4%	3.5%

more working-class communities. Almost half of the teacher workforce in both communities had 5 years or less experience in teaching kindergarten, and more than their bachelor's degree in education. Although there was a higher percentage of African American teachers in low-income neighborhoods, the majority of the teachers were Caucasian.

To conduct this analysis, we first compared the duration of observational times across groups to compare whether there were significant variations. As expected, there were no significant differences by SES group in observation time, t (42) = 1.42, p = .165 (see Table 6). This variable, therefore, was not included as a covariate in further analyses. Using MANOVA, we found no significant differences between groups on teachers' overall use of child-directed language, F (4, 39) = 2.72, p = .072. Planned univariate comparisons revealed no significant effects of lexical density, F (1, 42) = .73, p = .398 or the quantity of language, F (1, 42) = 2.27, p = .139. There were, however, significant differences between groups on sentence complexity (F [1, 42] = 7.03, p = .011, *partial* η^2 = .14), and lexical diversity (F [1, 42] = 4.84, p = .033, *partial* η^2 = .10). Children from lower income settings heard less varied words than those who attended kindergarten in more advantaged communities (see Table 6). Although the quantity of child-directed language in classrooms did not significantly differ, the quality of the language—both its complexity and variety—did differ, indicating a home/school continuity that may not be to these children's long-term advantage.

Parent and Teacher Language as Predictors of Child Outcomes

Our final question was whether these differences in parent and teacher child-directed language predicted children's language and school readiness outcomes. Table 8 provides descriptive statistics (means and standard deviations) for children's language and readiness skills from the beginning of the year to the end. To directly examine growth in children's outcomes over the year, we conducted a repeated measures MANOVA, with test type and time (pretest-to-posttest scores) as within-subject variables and neighborhood as a between-subjects variable. This analysis allowed us to account for children's pretest scores in order to examine differences in growth by neighborhood.

We found an overall significant effect of pre- to posttest gains, indicating significant growth across the school year for children in both SES groups (F [1, 52] = 10.93, p = .002 *partial* η^2 = .17). However, there was also a significant effect of neighborhood. Children in the working-class neighborhoods outperformed children in the low-income neighborhoods, starting at the beginning and end of the school year (F [1, 52] = 9.93, p = .003, *partial* η^2 = .16). Step-down tests showed that this difference was significant for each of the individual outcome measures (F [1, 52] = 9.84, p = .003, *partial* η^2 = .14) for Woodcock-Johnson Picture Vocabulary; F [1, 52] = 12.07, p = .001, *partial* η^2 = .18) for PPVT; F (1, 52) = 8.98, p = .004, *partial* η^2 = .13) for letter name knowledge; and F (1, 52) = 12.98, p = .001, *partial* η^2 = .18) for letter-sound knowledge. There was also a significant test type by time (pre- to posttest) by SES interaction, F (1, 50) = 3.73, p = .017. To explore this interaction, we calculated children's gains for each outcome measure over the year and examined differences by SES. There was a significant effect of SES on gains in expressive language (W-J), F (1, 52) = 8.08, p = .006, but no

Table 8
Descriptive Statistics (Means, Standard Deviations) of Children's Language and Literacy Outcomes

	End of preschool		Fall of kindergarten (Pre-test)		Spring of kindergarten (Post-test)	
Variable	Concentrated poverty	Borderline communities	Concentrated poverty	Borderline communities	Concentrated poverty	Borderline communities
WJIII-Picture Vocabulary[a]	99.25 (13.14)	101.87 (11.82)	103.45 (15.65)	111.09 (11.19)*	102.74 (9.71)	114.21 (15.26)***
PPVT-IV[b]	—	—	92.30 (15.14)	105.00 (14.19)***	99.93 (13.62)	110.11 (14.73)**
Letter names	—	—	.62 (.41)	.87 (.23)**	.82 (.27)	.96 (.11)**
Letter sounds	—	—	.28 (.34)	.48 (.37)*	.59 (.34)	.90 (.20)***
WJIII-Letter/Word identification	—	—	—	—	108.03 (10.13)	123.09 (17.66)***
WJIII-Word attack	—	—	—	—	76.48 (52.57)	94.27 (56.20)

[a] WJIII = Woodcock Johnson-III. [b] PPVT-IV = Peabody Picture Vocabulary Test-IV.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

significant effects on children's gains in letter name knowledge, (F [1, 52] = 3.32, $p = .074$. There were no significant effects of SES on gains in receptive language (PPVT), F (1, 52) = .06, $p = .801$, F (1, 52) = .02, $p = .878$. Taken together, these results suggest that even though both groups of children were learning, those from the working-class communities were soon outpacing their counterparts from low-income neighborhoods in expressive vocabulary.

We next correlated each of the parent and teacher variables with the child assessment measures (see Table 9). As expected, parents' sentence complexity was significantly correlated with children's receptive and expressive vocabulary. Lexical diversity was also significantly correlated with children's expressive, but not receptive, vocabulary. Of particular note, parents' reading comprehension skills were significantly correlated with all outcome measures including language characteristics as well as children's early reading skills. By contrast, teachers' sentence complexity was significantly correlated with children's letter-word identification and word attack skills, but not their receptive or expressive vocabulary knowledge. Other parent and teacher child-directed language variables (e.g., quantity; lexical diversity), were not significantly related to children's school readiness outcomes, and were therefore not included in further analyses or interpretation.

Next, we examined whether parent and teacher language variables predicted child outcomes using a series of hierarchical linear regressions. To create the most parsimonious model, we used each of the standardized child outcome measures at the end of the year as dependent variables in independent models. Parent sentence complexity, lexical diversity, and reading comprehension skill served as predictors at the first step for each model, teacher sentence complexity was entered at the second step, and finally, family SES was added at the third step. Because our predictor language variables were correlated, we first tested the assumption of collinearity. These results indicated that although these variables were correlated, they did not reach multicollinearity (parent sentence complexity, Tolerance = .42, VIF = 2.36, parent lexical diversity, Tolerance = .44, VIF = 2.26, parent WRAT, Tolerance = .87, VIF = 1.15, teacher sentence complexity, Tolerance = .95, VIF = 1.05, SES, Tolerance = .65, VIF = 1.54).

As shown in Tables 10–13, we found that home and school language variables contributed independently to the variance in children's school readiness outcomes. For children's receptive vocabulary knowledge (see Table 10), parent language variables significantly contributed to variance in children's PPVT scores as a whole. However, independently, only the contribution of parent reading skill (i.e., WRAT) was significant. Parent language variables also significantly contributed to children's expressive vocabulary ($p = .022$; see Table 11), although in this case, parent reading skill did not make a significant contribution.

By contrast, teacher language did not uniquely account for variance in children's receptive vocabulary knowledge ($p = .412$) or expressive language ($p = .505$; Tables 10-11).

Instead, teacher language significantly predicted letter-word identification scores ($p = .015$), explaining about 16% of the variance in children's scores (see Table 12). Teacher language was not a significant source of variance in children's WJIII-word attack scores ($p = .079$; see Table 13). Parent language did not significantly predict children's reading outcomes (letter-word identification: $p = .052$) or word attack scores: $p = .526$). Taken together, these results suggest that parent language and reading skills contributed to children's receptive and expressive language, whereas teacher language variables contributed primarily to children's early reading skills.

In order to determine whether these results were related to student-level SES, we added the individual family SES (a combination of family income, education, and H.O.M.E scores) as a predictive variable in the third step of the regression models. As shown in Tables 10-13, family SES accounted for variance in children's receptive vocabulary knowledge ($p = .007$), expressive vocabulary knowledge ($p = .003$), and letter-word identification skill ($p = .001$)

Table 9
Correlation Matrix for Parent language, Teacher Language, and Child Outcome Variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Parent language variables														
1. Duration of interaction	—													
2. Quantity of language	.29*	—												
3. Sentence complexity	.28*	.61**	—											
4. Lexical density	.21	.99**	.59**	—										
5. Lexical diversity	.33**	.89**	.72**	.88**	—									
6. WRAT4 scores	−.02	.04	.27*	.03	.23	—								
Teacher language variables														
7. Quantity of language	.14	−.06	−.33	−.06	−.17	−.33*	—							
8. Sentence complexity	−.004	−.10	−.01	−.10	−.09	.18	.35**	—						
9. Lexical density	.03	.02	−.16	−.05	.01	−.27	.65**	.31*	—					
10. Lexical diversity	.14	−.03	−.25	−.03	−.11	−.19	.88**	.43**	.59**	—				
Child outcome variables														
11. Receptive vocabulary	.06	−.02	.33*	−.03	.21	.46*	.07	.23	−.09	.10	—			
12. Expressive vocabulary	.12	.06	.39**	.06	.30*	.41*	.07	.21	−.09	.01	.81**	—		
13. Letter-word identification	.15	.09	.29*	.08	.26*	.32*	.09	.42**	.02	.14	.63**	.68**	—	
14. Word attack	−.01	−.15	.06	−.17	−.10	.27*	−.05	.37*	−.08	−.05	.12	.17	.02	—

Note. WRAT = Wide Range Achievement Test.
$^{*}p < .01$. $^{**}p < .001$.

Table 10
Hierarchical Regression Analysis of Parent and Teacher Variables Predicting Children's Receptive Vocabulary (PPVT-IV)

Step	ΔR^2	B^{a}	SE	t-value	p-value	sr^{2b}
Step 1	.32				.011	
Parent sentence complexity		9.07	520	1.74	.092	.10
Parent lexical diversity		−.10	.08	−1.32	.198	.06
Parent WRAT4		.485	.18	2.65	.013	.20
Step 2	.02				.412	
Parent sentence complexity		8.84	5.23	1.69	.102	.09
Parent lexical diversity		−.09	.08	−1.19	.243	.05
Parent WRAT4		.46	.19	2.43	.022	.18
Teacher sentence complexity		2.42	2.42	.83	.412	.02
Step 3	.16				.007	
Parent sentence complexity		4.26	4.9	.19	.867	.01
Parent lexical diversity		−.10	.07	−1.38	.180	.04
Parent WRAT4		.38	.17	2.23	.035	.09
Teacher sentence complexity		−.16	2.74	−.06	.954	<.001
Family SES		10.95	3.77	2.91	.007	.16

Note. WRAT4 = Wide Range Achievement Test; SES = socioeconomic status.
[a] Unstandardized regression coefficient. [b] Squared semi-partial correlation.

above and beyond the effects of parent or teacher language. However, family SES did not account for a significant amount of additional variance in children's word attack skills (p = .779).

Collectively, these results suggest that parent and teacher language relate to children's school readiness skills in significant, but somewhat distinct, ways. In particular, parent language seems to exert a stronger relationship on children's language development and their initial entry into reading. On the other hand, teacher language appears to contribute primarily to children's early decoding and reading skills. Student-level socioeconomic characteristics further contributed to these differences, suggesting the relationship among ecological factors of home, and school in these neighborhoods on children's school readiness skills.

Discussion

From an ecological perspective, this study posited that children's early language and reading opportunities may result from the interplay of environmental contexts. We specifically focused on the microsystems that support the most proximal processes: interactions between children and adults, as well as children's immediate environments of home and classroom contexts. At the same time, these proximal processes are themselves subject to environmental influence, and may systematically vary based on the characteristics of their surrounding context. Neighborhoods, which are organized or clustered in social settings, may exercise affordances or constraints and create "social milieus" that foster educational advantages or disadvantages. Consequently, in this study,

Table 11
Hierarchical Regression Analysis of Parent and Teacher Variables Predicting Children's Expressive Vocabulary (WJIII-PV)

Step	ΔR^2	B^{a}	SE	t-value	p-value	sr^{2b}
Step 1	.28				.022	
Parent sentence complexity		8.87	4.83	1.84	.076	.10
Parent lexical diversity		−.05	.07	−.67	.508	.015
Parent WRAT4		.31	.17	1.80	.083	.10
Step 2	.01				.505	
Parent sentence complexity		8.70	4.88	1.78	.085	.10
Parent lexical diversity		−.04	.07	−.57	.571	.01
Parent WRAT4		.28	.18	1.62	.117	.09
Teacher sentence complexity		1.83	2.72	.68	.505	.02
Step 3	.21				.003	
Parent sentence complexity		3.98	4.42	.90	.376	.02
Parent lexical diversity		−.04	.06	−.71	.486	.01
Parent WRAT4		.20	.15	1.32	.197	.03
Teacher sentence complexity		−.83	2.46	−.34	.738	.002
Family SES		11.28	3.39	3.33	.003	.21

Note. WRAT4 = Wide Range Achievement Test; SES = socioeconomic status.
[a] Unstandardized regression coefficient. [b] Squared semi-partial correlation.

Table 12

Hierarchical Regression Analysis of Parent and Teacher Variables Predicting Children's Letter-Word identification

Step	ΔR^2	B[a]	SE	t-value	p-value	sr^2[b]
Step 1	.23				.052	
Parent sentence complexity		11.10	6.26	1.78	.086	.10
Parent lexical diversity		−.095	.09	−1.03	.313	.03
Parent WRAT4		.37	.22	1.70	.101	.09
Step 2	.15				.015	
Parent sentence complexity		10.32	5.73	1.80	.082	.10
Parent lexical diversity		−.07	.085	−.79	.432	.02
Parent WRAT4		.27	.21	1.32	.197	.06
Teacher sentence complexity		8.25	3.20	2.59	.015	.16
Step 3	.22				.001	
Parent sentence complexity		4.21	4.95	.85	.402	.01
Parent lexical diversity		−.07	.07	−1.01	.324	.01
Parent WRAT4		.17	.17	.97	.341	.01
Teacher sentence complexity		4.80	2.76	1.74	.093	.04
Family SES		14.60	3.80	3.84	.001	.22

Note. WRAT4 = Wide Range Achievement Test; SES = socioeconomic status.
[a] Unstandardized regression coefficient. [b] Squared semi-partial correlation.

we have argued that we need to account for the embedded nature of these multiple contexts in our understanding of children's early development.

Therefore, the overall purpose of this study was to examine the influence of environmental supports as young children transitioned from preschool to kindergarten. We focused our inquiry on children from two distinctive areas. The first, identified as concentrated poverty, included children who had attended a county-wide Head Start preschool program and lived in an area of extreme urban blight, high unemployment, and child poverty. The second, identified as working-class, included children who had attended a state-funded preschool program and lived in small borderline neighborhoods located at the periphery of the high-poverty area. Implicit in this inquiry was the role that environment—home and school plays in the development of early literacy.

To better understand this larger context, we attempted to look beyond early childhood programs for an explanation. From our perspective, too often we have focused on what happens within various early childhood programs rather than on the environmental supports that surround them. In this respect, our goal was not to compare the early childhood programs children had attended. In fact, based on outside evaluations, both the Head Start program and the statewide pre-K program used similar standards and had received high marks (Head Start Bureau, 2001; Xiang & Schweinhart, 2002). Instead, our goal was to examine children's experience as they transitioned to the kindergarten classroom in these communities. Given the well-documented role that adults' use of child-directed language plays in children's language and literacy development, we focused on the conceptual content of the linguistic interactions

Table 13

Hierarchical Regression Analysis of Parent and Teacher Variables Predicting Children's Word Attack (WJIII-WA) Scores

Step	ΔR^2	B[a]	SE	t-value	p-value	sr^2[b]
Step 1	.08				.526	
Parent sentence complexity		23.54	20.16	1.17	.253	.05
Parent lexical diversity		−.35	.30	−1.15	.259	.05
Parent WRAT4		.47	.71	.67	.511	.02
Step 2	.10				.079	
Parent sentence complexity		21.98	19.39	1.13	.267	.05
Parent lexical diversity		−.29	.29	−1.00	.328	.04
Parent WRAT4		.19	.70	.26	.794	.002
Teacher sentence complexity		20.65	11.32	1.82	.079	.11
Step 3	.003				.779	
Parent sentence complexity		20.09	20.82	.97	.344	.03
Parent lexical diversity		−.29	.29	−.98	.335	.03
Parent WRAT4		.15	.72	.21	.837	.001
Teacher sentence complexity		19.66	12.04	1.63	.114	.08
Family SES		4.54	16.02	.28	.779	.003

Note. WRAT4 = Wide Range Achievement Test; SES = socioeconomic status.
[a] Unstandardized regression coefficient. [b] Squared semi-partial correlation.

in both home and school contexts, as well as their subsequent effects on children's language and early reading skills.

Our results suggest that no matter the strength of the early boost children receive from their preschool program, differences in subsequent environmental influences can either support or undermine any early advantage. Children who lived in a neighborhood of concentrated poverty had fewer supports for school-based language and early literacy development than those in working-class communities. Families who themselves were likely to have attended low-quality schools in these poor neighborhoods (Sharkey, 2013) had less education and lower reading comprehension skills than their working-class peers. These findings further support previous studies highlighting the structural inequality of early childhood educational contexts and how they may account for children's outcomes over time (Hoff, 2013; McCoy, Connors, Morris, Yoshikawa, & Friedman-Krauss, 2015).

Children's transition to elementary school appeared to follow an unfortunate, but similar pattern. Given that school placements are almost entirely dependent on residential location (Lareau & Goyette, 2014), children in these poor communities went on to attend more economically segregated schools as compared to those in the working-class communities, which were typically more economically diverse. These children from low-income neighborhoods were also more likely to attend a school that had failed to meet adequate yearly progress (21% according to NCLB) that had shorter hours (i.e., half-day kindergarten vs. full-day), with teachers of less experience. In contrast to studies that have examined wide economic disparities, even in these neighborhood contexts that were not all that dissimilar, there were differences in the composition of the schools. Tragically, the children who might need the greater opportunity to learn appeared to be the least likely to get it.

Instead, these children from the poorest community attended kindergartens characterized by more limited language opportunities. In fact, one could argue that many teachers oversimplified their language for these children. They used simpler sentences of more limited construction than the teachers of children who attended classes in borderline communities. There was also less variation and fewer unique word types in their language. In short, these children received a "double dose of disadvantage," as they transition from home to school contexts.

These results may have serious implications. Huttenlocher, Vasilyeva, Cymerman, and Levine (2002), for example, have demonstrated a positive relationship between the complexity and variety of teachers' language input and children's language growth. Unfortunately, in our case this thesis remained untested since such language features were not particularly high for either group. In both communities, teachers' child-directed language did not contribute to receptive or expressive language. However, teachers' child-directed language did contribute to children's letter-word identification skills. These results might reflect teachers' greater attention to alphabetic skills, or the process of lexical restructuring (Metsala, 1999). For example, studies (Metsala, 1997; Walley, 1993) have shown that as children experience more complex sentence structures, they begin to change from storing words holistically to representing words in more discrete phonological units of sound. Clearly, further research is needed to disentangle this relationship.

What was most apparent in our investigation of this transitional year was not necessarily a *decline* in language and literacy development for children who lived in the poorest community; rather, it was their limited *growth*. For example, expressive language skills were essentially stable for these children as compared to the previous year's assessment. Nevertheless, although their skills grew throughout the kindergarten year, they grew at a slower pace than their neighborhood counterparts. This pattern stands in stark contrast to the steady, albeit modest, gains in expressive language demonstrated by children living in the working-class communities. In other words, while nominally still "average" in expressive language skills, there was now more than a full standard deviation separating these two groups by the end of the kindergarten year.

Given this discrepancy in school-based language and knowledge-building supports, the lack of a sustained pattern of growth in school readiness skills over time is hardly surprising. Although some researchers have posited that the problem lay in the wide variation in program quality in Head Start and other targeted preschool programs across sites (Ramey et al., 2000), rarely have the subsequent differences in the educational contexts (specifically the home and school contexts in neighborhoods to which our poorest children are exposed to) prominently feature in our policy discussions. In this study, we compared children who came from a neighborhood of concentrated poverty, identified as above 40% of children and family who are poor, to children who came from borderline communities showing signs of economic distress, with poverty rates ranging from 20–40%. If such significant differentials in the home and school contexts were so clearly evident in these communities, one can only imagine the differences in opportunities between these children and other children who live in more affluent circumstances. Yet previous research on the sustained effects of early education programs have, by and large, assumed that the out-of-school experiences of program participants are equivalent to those of their comparison-group peers (Lipsey, Hofer, Dong, Farran, & Bilbrey, 2013; Xiang & Schweinhart, 2002). This study demonstrates that this is simply not the case.

This study highlights that home and school contexts themselves reside in a specific context, the neighborhood, and this spatial concentration may influence stratification and social isolation. William Julius Wilson (1987), in his classic sociological analysis of demographic changes in Chicago, showed that we cannot explain young people's behavior simply by understanding individual disadvantage, but must account for neighborhood influences as well. In neighborhoods of concentrated poverty, individual families must cope not only with their own poverty, but also with the economic deprivation of the many families who live nearby. This spatial concentration can act to magnify economic disadvantage and exacerbate its effects (McCoy et al., 2015). Sharkey (2013), for example, found that the effects of neighborhood disadvantage during childhood continued to have strong impacts through adulthood. Moreover, a recent study by Chetty, Hendren, and Katz (2015) showed the potential benefits when low-SES families move to a good neighborhood. Through a reanalysis of the *Moving to Opportunity* study in which parents through lottery were given vouchers to move to better neighborhoods, they found that the earlier a family moved to a good neighborhood, the better the children's long-term outcomes. In particular, children whose families who moved before they were 8 years old went on to earn 31%

more than children whose families did not win the lottery. This research, in short, suggests that neighborhood matters and can have a powerful influence on nurturing success (or failure).

There are alternative explanations, however, to our finding that these multiple contexts exerted a powerful influence and contributed to children's school readiness. Some might argue that differentials in children's language and school readiness skills reflected a "sleeper effect" (Duncan & Murnane, 2011), with the state targeted pre-K program demonstrating longer-term sustainability than the Head Start program. However, a recent analysis of teachers' enactment of instruction using teacher logs in 14 state-funded classroom programs showed a striking similarity in programmatic features as in the Head Start program (Herald, 2015). Others might legitimately argue that the measures of school readiness in the present study were relatively limited. Given the time constraints and logistics involved in assessing children, their parents, and their teachers, it was necessary to include a smaller number of measures. Further work should include a more comprehensive battery of school readiness assessments. Additionally, our sample size was relatively small. Therefore, we cannot assume the generalizability of our findings. Subsequent research including larger samples comparing neighborhoods of concentrated poverty, borderline communities, and middle-income communities might better gauge the out-of-school educational opportunities and supports needed to improve and sustain children's school readiness.

Furthermore, one might consider the close relationship between our contexts as constituting a confounding effect. It is true that our study could not disentangle the independent effects of family, school and neighborhood on child outcomes, and in this respect, it represents an important limitation. On the other hand, however, we suggest that a central point of our argument was to describe and better understand how these contexts are embedded in one another, and how a complex web of social contexts may shape children's development. Consistent with Bronfenbrenner and Morris's (1998) ecological model of human development, it emphasizes the role that different environmental systems may have both directly and indirectly on children's early learning experiences.

It may also suggest a more expansive approach to intervention for the future than one that focuses on the word or language "gap." Instead, it may argue that the structural inequalities of these environments call for a very different research enterprise. For example, Bryk and his colleagues (Bryk, Gomez, Grunow, & LeMhieu, 2015) have introduced the idea of networked improvement communities that build on creating purposeful collective actions within and among community members, using the iterative process of improvement science to guide, revise and fine-tune interventions that meet the needs of its citizens. In this respect, it begins to build on the strengths of the community, and not its deficits. It also begins to take into account the multiple environments that influence children's learning, and the ways in which various institutions might better collaborate to accelerate continuous improvement.

At the same time, the present study offers new and additional support for the important role that parents' own reading comprehension may contribute to children's development. Studies have shown that mothers' education predicts children's early literacy skills (Denton et al., 2003; Mol & Neuman, 2014); however, significant relationships between parents' reading comprehension abilities and their children's own early language and literacy skills have been less well documented. Replicating these results with a larger sample could help to explain the substantial differences in the early childhood read-aloud practices documented between low- and middle-income families (Common Sense Media, 2013). Although parents may involve their children in many other literacy-related activities such as storytelling (Heath, 1983), they might seek alternatives to book-reading. If parents themselves struggle with reading, it is unlikely that they will regard reading aloud as an enjoyable task to engage in with their children. Therefore, if such a connection (e.g., poor reading skills, limited reading aloud to children) is shown in subsequent research, it could help to explain the unique effects of being read to early on. Many researchers (Cunningham & Zibulsky, 2014; Cunningham & Stanovich, 1998) have suggested that reading volume, even more than oral language, is the prime contributor to individual differences in children's vocabularies.

Previous studies (e.g., Aram & Levin, 2004) have shown that parent supports, namely oral and written language opportunities in the home, significantly contribute to children's language and literacy development. Studies have also shown that there are striking differences in these supports in the United States for families of different socioeconomic status compared to other industrialized nations (Bradbury, Corak, Waldfogel, & Washbrook, 2015). Our study has substantiated these findings. Other studies (e.g., Reardon, 2011) have reported strong associations between children's SES status and the quality of the schools they attend, representing a de facto segregation of schools by family economic condition. We similarly found significant student-level SES differences in the schools and classrooms children attend. However, unique to this study, was a more inclusive analysis looking at both of the settings within communities that have become increasingly segregated, not exclusively by race, but by class. In doing so, our study highlights that the playing field for entering kindergartners is hardly equal; rather, the quality of one's educational opportunities is highly dependent on the streets where you live. Consequently, improving children's chances of school success will require a far more comprehensive approach than a 1-year preschool program, no matter how successful it may be.

References

Aram, D., & Levin, I. (2004). The role of maternal mediation of writing to kindergarteners in promoting literacy in school: A longitudinal perspective. *Reading and Writing, 17,* 387–409. http://dx.doi.org/10.1023/B:READ.0000032665.14437.e0

Bailey, D., Duncan, G., Odgers, C., & Yu, W. (2015). *Persistence and fadeout in the impacts of child and adolescent interventions.* Brisbane, Australia: Institute for Social Science Research, the University of Queensland.

Beck, I., & McKeown, M. (2007). Increasing young low-income children's oral vocabulary repertoires through rich and focused instruction. *The Elementary School Journal, 107,* 251–271. http://dx.doi.org/10.1086/511706

Bischoff, K., & Reardon, S. F. (2014). Residential segregation by income, 1970–2009. In J. R. Logan (Ed.), *Diversity and disparities: America enters a new century* (pp. 208–233). New York, NY: Russell Sage Foundation.

Bradbury, B., Corak, M., Waldfogel, J., & Washbrook, E. (2015). *Too many children left behind.* New York, NY: Russell Sage Foundation.

Bradley, R. H., Caldwell, B. M., & Corwyn, R. F. (2003). The child care HOME inventories: Assessing the quality of family child care homes.

Early Childhood Research Quarterly, 18, 294–309. http://dx.doi.org/10.1016/S0885-2006(03)00041-3

Brandone, A. C., Cimpian, A., Leslie, S. J., & Gelman, S. A. (2012). Do lions have manes? For children, generics are about kinds rather than quantities. *Child Development, 83,* 423–433.

Bronfenbrenner, U., & Morris, P. (1998). The ecology of developmental processes. In W. Damon (Series Ed.) & R. M. Lerner (Vol. Ed.), *Handbook of child psychology: Vol. I: Theoretical models of human development* (5th ed., pp. 2–1028). New York, NY: Wiley.

Brooks-Gunn, J., Duncan, G., & Aber, J. L. (Eds.). (1997). *Neighborhood poverty.* New York, NY: Russell Sage Foundation.

Brown, P., Roediger, H., & McDaniel, M. (2014). *Make it stick.* Cambridge, MA: Harvard University Press. http://dx.doi.org/10.4159/9780674419377

Brown, R. (1973). *A first language: The early stages.* Cambridge, MA: Harvard University Press. http://dx.doi.org/10.4159/harvard.9780674732469

Bryk, A., Gomez, L., Grunow, A., & LeMahieu, P. (2015). *Learning to improve: How America's schools can get better at getting better.* Cambridge, MA: Harvard Education Press.

Buckingham, J., Wheldall, K., & Beaman-Wheldall, R. (2013). Why poor children are more likely to become poor readers: The school years. *Australian Journal of Education, 57,* 190–213. http://dx.doi.org/10.1177/0004944113495500

Chetty, R., Hendren, N., & Katz, L. F. (2015). *The long-term effects of exposure to better neighborhoods: New evidence from the Moving to Opportunity experiment.* Cambridge, MA: National Bureau of Economic Research, Harvard University. http://dx.doi.org/10.3386/w21156

Chomsky, N. (1968). *Language and mind.* New York, NY: Harcourt, Brace, Jovanovich. http://dx.doi.org/10.1037/e400082009-004

Common Sense Media. (2013). *Zero to eight: Children's media use in America 2013.* Washington, DC: Common Sense Media.

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33,* 934–945. http://dx.doi.org/10.1037/0012-1649.33.6.934

Cunningham, A. E., & Stanovich, K. (1998). What reading does for the mind. *American Educator, 22,* 8–15.

Cunningham, A., & Zibulsky, J. (2014). *Book smart.* New York, NY: Oxford University Press.

Denton, K., West, J., & Waltston, J. (2003). *Reading: Young children's achievement and classroom experiences: Findings from The Condition of Education 2003.* Washington, DC: National Center for Educational Statistics.

Dickinson, D., Hofer, K., Barnes, E., & Grifenhagen, J. (2014). Examining teachers' language in Head Start classrooms from a Systemic Linguistics Approach. *Early Childhood Research Quarterly, 29,* 231–244. http://dx.doi.org/10.1016/j.ecresq.2014.02.006

Dickinson, D. K., & Porche, M. V. (2011). Relation between language experiences in preschool classrooms and children's kindergarten and fourth-grade language and reading abilities. *Child Development, 82,* 870–886. http://dx.doi.org/10.1111/j.1467-8624.2011.01576.x

Duncan, G., & Murnane, R. (2011). *Whither opportunity?* New York, NY: Russell Sage Foundation.

Dunn, L., & Dunn, D. (2007). *Peabody Picture Vocabulary Test* (4th ed.). Bloomington, MN: Pearson Education, Inc.

Dupere, V., Leventhal, T., Crosnoe, R., & Dion, E. (2010). Understanding the positive role of neighborhood socioeconomic advantage in achievement: The contribution of the home, child care, and school environments. *Developmental Psychology, 46,* 1227–1244. http://dx.doi.org/10.1037/a0020211

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science, 16,* 234–248. http://dx.doi.org/10.1111/desc.12019

Galbraith, S., Daniel, J. A., & Vissel, B. (2010). A study of clustered data and approaches to its analysis. *The Journal of Neuroscience, 30,* 10601–10608. http://dx.doi.org/10.1523/JNEUROSCI.0362-10.2010

Gelman, S., & Kalish, C. (2006). Conceptual development. In D. Kuhn, R. S. Siegler, W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology: Vol. 2: Cognition, perception, and language* (6th ed. pp. 687–733). New York, NY: Wiley.

Gelman, S. A., & Raman, L. (2003). Preschool children use linguistic form class and pragmatic cues to interpret generics. *Child Development, 74,* 308–325. http://dx.doi.org/10.1111/1467-8624.00537

Harris, J., Golinkoff, R. M., & Hirsh-Pasek, K. (2011). Lessons from the crib for the classroom: How children really learn vocabulary. In S. B. Neuman & D. Dickinson (Eds.), *Handbook of early literacy research* (Vol. III, pp. 49–65). New York, NY: Guilford Press.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore, MD: Brookes.

Head Start Bureau. (2001). Head Start Child Outcomes Framework. *National Head Start Bulletin, 70,* 44–50. Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/tta-system/pd/pds/Cultivating%20a%20Learning%20Organization/ScreenAssess.pdf

Heath, S. B. (1983). *Ways with words: Language, life, and work in communities and classrooms.* New York, NY: Cambridge University Press.

Herald, L. (2015). *At-risk preschool children's exposure to amounts and types of literacy and science instruction* (Unpublished Dissertation). University of Michigan, Ann Arbor, MI.

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review, 26,* 55–88. http://dx.doi.org/10.1016/j.dr.2005.11.002

Hoff, E. (2013). Interpreting the early language trajectories of children from low-SES and language minority homes: Implications for closing achievement gaps. *Developmental Psychology, 49,* 4–14. http://dx.doi.org/10.1037/a0027238

Hoff, E., Rumiche, R., Burridge, A., Ribot, K. M., & Welsh, S. N. (2014). Expressive vocabulary development in children from bilingual and monolingual homes; A longitudinal study from two to four years. *Early Childhood Research Quarterly, 29,* 433–444. http://dx.doi.org/10.1016/j.ecresq.2014.04.012

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology, 45,* 337–374. http://dx.doi.org/10.1016/S0010-0285(02)00500-5

Jargowsky, P. A. (2014). *Concentration of poverty in the new millennium.* Washington, DC: Century Fund.

Landry, S. H., Smith, K. E., Swank, P. R., Assel, M. A., & Vellet, S. (2001). Does early responsive parenting have a special importance for children's development or is consistency across early childhood necessary? *Developmental Psychology, 37,* 387–403. http://dx.doi.org/10.1037/0012-1649.37.3.387

Lareau, A., & Goyette, K. (Eds.). (2014). *Choosing homes, choosing schools.* New York, NY: Russell Sage Foundation.

LaRusso, M., Kim, H. Y., Selman, R., Uccelli, P., Dawson, T., Jones, S., . . . Snow, C. (2016). Contributions of academic language, perspective taking, and complex reasoning to deep reading comprehension. *Journal of Research on Educational Effectiveness, 9,* 201–222. http://dx.doi.org/10.1080/19345747.2015.1116035

Lipsey, M. W., Hofer, K. G., Dong, N., Farran, D. C., & Bilbrey, C. (2013). *Evaluation of the Tennessee Voluntary Prekindergarten Program: Kindergarten and first grade follow-up results from the randomized control design.* Nashville, TN: Peabody Research Institute, Vanderbilt University.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Erlbaum.

Massey, D. S. (2007). *Categorically unequal: The American stratification system.* New York, NY: Russell Sage Foundation.

Mayer, S. E. (1998). *What money can't buy: Family income and children's life chances*. Cambridge, MA: Harvard University Press.

McCoach, D. B., & Adelson, J. (2010). Dealing with dependence (Part 1): Understanding the effects of clustered data. *Gifted Child Quarterly, 54,* 152–155. http://dx.doi.org/10.1177/0016986210363076

McCoy, D. C., Connors, M. C., Morris, P. A., Yoshikawa, H., & Friedman-Krauss, A. H. (2015). Neighborhood economic disadvantage and children's cognitive and social-emotional development: Exploring Head Start classroom quality as a mediating mechanism. *Early Childhood Research Quarterly, 32,* 150–159. http://dx.doi.org/10.1016/j.ecresq.2015.04.003

Metsala, J. (1997). Spoken word recognition in reading disabled children. *Journal of Educational Psychology, 89,* 159–169. http://dx.doi.org/10.1037//0022-0663.89.1.159

Metsala, J. (1999). Young children's phonological awareness and nonword repetition as a function of vocabulary development. *Journal of Educational Psychology, 91,* 3–19. http://dx.doi.org/10.1037/0022-0663.91.1.3

Michigan Great Start Readiness Program. (2012). Retrieved from http://www.michigan.gov/mde/0,4615,7-140-63533_50451---,00.html

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin, 137,* 267–296. http://dx.doi.org/10.1037/a0021890

Mol, S., & Neuman, S. B. (2014). Sharing information books with kindergartners: The role of parents' extra-textual talk and socio-economic status. *Early Childhood Research Quarterly, 29,* 399–410. http://dx.doi.org/10.1016/j.ecresq.2014.04.001

Neuman, S. B., & Carta, J. (2011). Advancing the measurement of quality for early childhood programs that support early language and literacy development. In M. Zaslow, I. Martinez-Beck, K. Tout, & K. Halle (Eds.), *Quality measurement in early childhood settings* (pp. 220–255). Baltimore, MD: Brookes.

Neuman, S. B., & Celano, D. (2012). *Giving our children a fighting chance: Affluence, literacy, and the development of information capital.* New York, NY: Teachers College Press.

Neuman, S. B., Newman, E., & Dwyer, J. (2011). Educational effects of a vocabulary intervention on preschoolers' word knowledge and conceptual development: A cluster randomized trial. *Reading Research Quarterly, 46,* 249–272.

NICHD Early Child Care Research Network. (1999). Child care and mother-child interaction in the first three years of life. *Developmental Psychology, 35,* 1399–1413. http://dx.doi.org/10.1037/0012-1649.35.6.1399

Oppenheim, J. (1996). *Have you seen bugs?* New York, NY: Scholastic.

Puma, M., Bell, S., Cook, R., & Heid, C. (2010). *Head Start Impact Study: Final report*. Washington, DC: Administration for Children and Families, U. S. Department of Health and Human Services.

Ramey, S. L., Ramey, C. T., Phillips, M. M., Lanzi, R. G., Brezausek, C., Katholi, C. R., & Snyder, S. (2000). *Head Start children's entry into public school: A report on the National Head Start/Public School Early Childhood Transition Demonstration Study*. Washington, DC: Administration on Children, Youth, and Families, Department of Health and Human Services.

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity? Rising inequality, schools, and children's life chances* (pp. 91–115). New York, NY: Russell Sage Foundation.

Rowe, M. L., Raudenbush, S. W., & Goldin-Meadow, S. (2012). The pace of vocabulary growth helps predict later vocabulary skill. *Child Development, 83,* 508–525.

Shanahan, T., Cunningham, A., Escamilla, K. C., Fischel, J., Landry, S., Lonigan, C. J., . . . Strickland, D. (2008). *Developing early literacy: Report of the National Early Literacy Panel*. Washington, DC: National Institute for Literacy.

Sharkey, P. (2013). *Stuck in place: Urban neighborhoods and the end of progress toward racial equality*. Chicago, IL: University of Chicago Press. http://dx.doi.org/10.7208/chicago/9780226924267.001.0001

Sigel, I. E. (1982). The relationship between parental distancing strategies and the child's cognitive behavior. In L. M. Laosa & I. E. Sigel (Eds.), *Families as learning environments for children* (pp. 47–86). New York, NY: Plenum Press. http://dx.doi.org/10.1007/978-1-4684-4172-7_2

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75,* 417–453. http://dx.doi.org/10.3102/00346543075003417

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360–407. http://dx.doi.org/10.1598/RRQ.21.4.1

Totsika, V., & Sylva, K. (2004). The Home Observation for Measurement of the Environment (HOME) revisited. *Child and Adolescent Mental Health, 9,* 25–35. http://dx.doi.org/10.1046/j.1475-357X.2003.00073.x

van Tuijl, C., & Leseman, P. (2004). Improving mother-child interaction in low-income Turkish-Dutch families: A study of mechanisms mediating improvements resulting from participating in a home-based preschool intervention program. *Infant and Child Development, 13,* 323–340. http://dx.doi.org/10.1002/icd.363

Vasilyeva, M., Waterfall, H., & Huttenlocher, J. (2008). Emergence of syntax: Commonalities and differences across children. *Developmental Science, 11,* 84–97.

Walley, A. (1993). The role of vocabulary development in children's spoken word recognition and segmentation ability. *Developmental Review, 13,* 286–350. http://dx.doi.org/10.1006/drev.1993.1015

Weizman, Z. O., & Snow, C. E. (2001). Lexical input as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental Psychology, 37,* 265–279. http://dx.doi.org/10.1037/0012-1649.37.2.265

Wilkinson, G. S., & Robertson, G. J. (2006). *Wide range achievement test* (4th ed.). Lutz, FL: Psychological Assessment Resources.

Wilson, W. J. (1987). *The truly disadvantaged: The inner city, the underclass, and public policy*. Chicago, IL: University of Chicago Press.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2007). *Woodcock-Johnson Psychoeducational Battery* (3rd ed.). Itasca, IL: Riverside Publishing.

Wright, T. S., & Neuman, S. B. (2014). Paucity and disparity in kindergarten oral vocabulary instruction. *Journal of Literacy Research, 46,* 330–357. http://dx.doi.org/10.1177/1086296X14551474

Xiang, Z., & Schweinhart, L. J. (2002). The Michigan School Readiness Program evaluation through age ten. *Journal of At-Risk Issues, 8,* 17–23. Retrieved from http://ndpc-web.clemson.edu/journals/journal-risk-issues-online-issues

Received July 7, 2016
Revision received February 21, 2017
Accepted February 24, 2017 ■

Journal of Educational Psychology
2018, Vol. 110, No. 1, 119–132

0022-0663/18/$12.00 http://dx.doi.org/10.1037/edu0000192

The Role of Authoritative and Authoritarian Parenting in the Early Academic Achievement of Latino Students

Yeonwoo Kim and Esther J. Calzada
The University of Texas at Austin

R. Gabriela Barajas-Gonzalez, Keng-Yen Huang, Laurie M. Brotman, Ashley Castro, and Catherine Pichardo
New York University School of Medicine

Early academic achievement has been shown to predict high school completion, but there have been few studies of the predictors of early academic success focused on Latino students. Using longitudinal data from 750 Mexican and Dominican American families, this study examined a cultural model of parenting and early academic achievement. While Latino students were achieving in the average range as a whole, certain subgroups (e.g., Dominicans, boys) were at higher risk for underachievement. Results highlighted the protective role of authoritative parenting, which was associated with academic and social-emotional school readiness, both of which predicted higher achievement at the end of first grade. The role of *respeto* and authoritarian parenting practices in academic achievement at first grade differed between Mexican and Dominican American families. Findings advance understanding of early achievement and parenting among Latino families from a cultural perspective.

Educational Impact and Implications Statement

Certain subgroups of Latino students (e.g., Dominican-origin, boys) are at greater risk for underachievement than their peers (e.g., Mexican-origin, girls). Authoritative parenting practices are a protective factor for academic achievement in first grade. Efforts to support Latina mothers in how they socialize and interact with their young children may have positive effects on their child's later academic achievement.

Keywords: academic achievement, early childhood, Latino students, parenting

According to the life course perspective of high school dropout (Alexander, Entwisle, & Kabbani, 2001), the successful completion of high school is determined by the interplay of individual child, family, and school factors that unfold well before students enter high school. In support of this model, a host of indicators in first grade, including behavior problems, school performance, grade retention, parent involvement, and family stressors have been shown to predict dropout (e.g., Brooks-Gunn, Guo, & Furstenberg, 1993; Cairns & Cairns, 1994; Ensminger & Slusarcick, 1992; Garnier, Stein, & Jacobs, 1997; Haveman & Wolfe, 1994; Roderick, 1994). In fact, even before they enter formal schooling, children have had a wealth of experiences that set the stage for the ways in which they will transition to and navigate through school (Finn, Gerber, & Boyd-Zaharias, 2005; Garnier et al., 1997). In a prospective, longitudinal study of children from 6 months through 19 years, Egeland and colleagues (Jimerson, Egeland, Sroufe, & Carlson, 2000) examined characteristics prior to school entry and throughout children's schooling as predictors of high school dropout. In their sample of non-Latino white and black families living in poverty, the home environment and quality of early parenting uniquely predicted dropout status, even when accounting for predictors measured at later timepoints. Despite the policy and practice implications of these findings, it is not known whether such findings apply to Latino student populations, a gap that is particularly notable in light of the disproportionate risk for school dropout experienced by Latinos (Aud, Fox, & KewalRamani, 2010). To address this limitation in the literature, the present study examined parenting as a predictor of early academic achievement in a sample of Latino students and their mothers.

Latino Parenting

Authoritative and Authoritarian Parenting Practices

A vast literature shows the strong and direct effects that parenting has on child development. Traditional parenting theory emphasizes responsiveness (i.e., warmth and nurturance) and demandingness (i.e., behavior management practices) as key dimensions underlying parenting styles. These global dimensions

This article was published Online First March 30, 2017.

Yeonwoo Kim and Esther J. Calzada, School of Social Work, The University of Texas at Austin; R. Gabriela Barajas-Gonzalez, Keng-Yen Huang, Laurie M. Brotman, Ashley Castro, and Catherine Pichardo, The Department of Population Health, New York University School of Medicine.

Correspondence concerning this article should be addressed to Esther J. Calzada, School of Social Work, The University of Texas at Austin, 1925 San Jacinto Boulevard, Austin, TX 78712. E-mail: esther.calzada@austin.utexas.edu

of parenting appear to be cross-culturally robust in that parents from all cultures place demands on and respond to the needs of their children, but some research suggests that different levels of demandingness and responsiveness may be seen across ethnic groups (e.g., Cardona, Nicholson, & Fox, 2000). More importantly, relations between parenting and child outcomes appear to be moderated by ethnicity (Hill, Bush, & Roosa, 2003; Steinberg, Lamborn, Dornbusch, & Darling, 1992).

For non-Latino White children, authoritative parenting practices (i.e., high responsiveness and demandingness) seem to promote, and authoritarian practices (i.e., low responsiveness and high demandingness) to hinder the developmental competencies needed to achieve academically. Among Latinos, the literature is inconclusive regarding the nature and influence of parenting practices according to a number of studies with largely Mexican-origin parents, the largest Latino subgroup in the U.S. (note, though, that studies rarely disaggregate findings by country of origin). Some studies suggest that Latino parents are highly authoritarian, and that their elementary- and high school–age children may not be impacted negatively by this parenting approach (Hillstrom, 2009; Moon, Kang, & An, 2009; Pong, Hao, & Gardner, 2005), whereas studies with Mexican- and Dominican-origin preschoolers find negative effects of authoritarian parenting (Calzada, Barajas-Gonzalez, Huang, & Brotman, 2015; Calzada, Huang, Anicama, Fernandez, & Brotman, 2012). These inconsistencies may reflect the differential impact of parenting practices across developmental stages; during early childhood, it may be expected that authoritarian parenting is especially maladaptive. For example, it seems likely that high levels of parental control may confer risk when children transition to formal schooling if controlling parenting inhibits the autonomy children need to be successful in adapting to a classroom setting. At the same time, although high levels of authoritarian control may be protective among adolescents who are vulnerable to negative peer and neighborhood influences, high control may not be necessary with preschoolers who have limited exposure to such risk factors.

Latino Cultural Socialization

Like all family processes, parenting is rooted in culture (i.e., the shared values, beliefs and experiences that determine a group's behavioral norms; Guarnaccia & Rodriguez, 1996). In the present study, we consider a cultural framework (Calzada, Fernandez, & Cortes, 2010; see Figure 1) that expands the focus of parenting research to emphasize cultural socialization as the impetus for the intentional use of authoritarian or authoritative parenting practices. Cultural socialization is the process through which parents transmit cultural values, beliefs, traditions, and behavioral norms to their children (Hughes et al., 2006). For Latino parents, adherence to the cultural value of *respeto*, which emphasizes obedience and deference to adults (Calzada, 2010; Calzada et al., 2010), appears to go hand-in-hand with an authoritarian parenting style. Depending in part on their acculturative status (i.e., acculturation, or adaptation to mainstream culture; and enculturation, or maintenance of one's culture of origin), Latino parents may emphasize *respeto* to varying degrees (Gonzales et al., 2008). More acculturated Latina mothers may be more likely to embrace the mainstream U.S. cultural value of independence, socializing their child to be assertive and autonomous, even during early childhood. The value of independence appears to go hand-in-hand with authoritative parenting practices. Thus, according to this framework (see Figure 1), mothers' acculturative status is expected to influence socialization messages of *respeto* or independence and the authoritarian or authoritative parenting practices that reinforce them; both, in turn, are expected to influence child developmental outcomes.

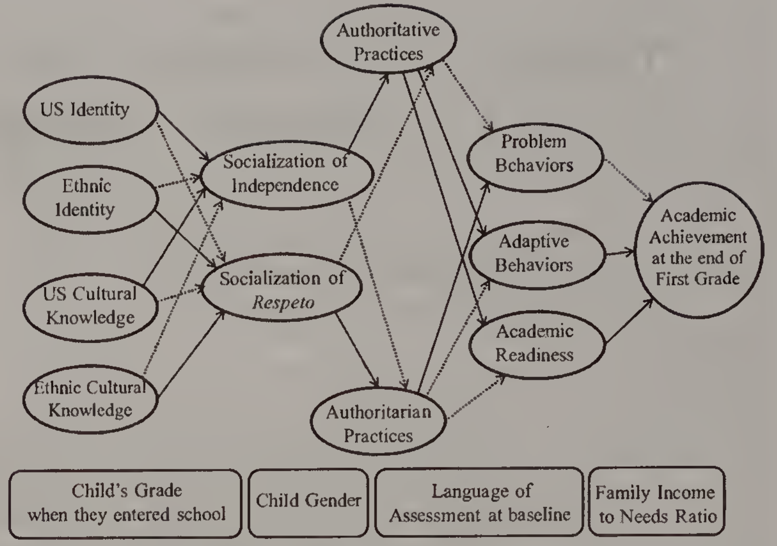

Authoritative Practices
US Identity
Ethnic Identity
US Cultural Knowledge
Ethnic Cultural Knowledge
Socialization of Independence
Socialization of *Respeto*
Authoritarian Practices
Problem Behaviors
Adaptive Behaviors
Academic Readiness
Academic Achievement at the end of First Grade
Child's Grade when they entered school
Child Gender
Language of Assessment at baseline
Family Income to Needs Ratio

Figure 1. Cultural framework of Latino parenting and student academic achievement. Straight lines indicate positive hypothesized associations, and dashed lines indicate negative hypothesized associations.

No studies have examined the relation of cultural socialization to children's early academic functioning but theoretically, *respeto* may make children more "teachable" in the classroom in that they respect authority and conform to rules (Harwood, 1992). Alternately, it may leave children ill-prepared for schooling in the U.S. educational system because they have been less encouraged to question, problem-solve or negotiate. Unexpectedly, in a correlational study of Latino preschoolers, Calzada and colleagues (2012) found *respeto* to be negatively associated with socioemotional functioning (a component of school readiness) among Mexican-origin (MA) and Dominican-origin (DA) students, mediated by authoritarian parenting practices, a finding that was later replicated using longitudinal data (Calzada et al., 2015). In contrast, independence positively influenced socioemotional functioning, though only among MA children, mediated by authoritative parenting practices. The authors speculated that the influence of cultural socialization could vary across MA and DA groups because they occur within unique ecological circumstances (e.g., an ethnic enclave vs. a diverse neighborhood; see Calzada et al., 2012 for further discussion). These studies not only highlight ethnic subgroup differences but also demonstrate the value of studying socialization messages in relation to the developmental competencies that Latino children need to be successful in school.

The Role of School Readiness

School readiness, or the extent to which children enter school with the foundational academic and social-emotional competencies that allows them to learn, predicts later achievement (Duncan et al., 2007). The academic domain of school readiness includes early language, understanding of concepts, and motor skills that serve as building blocks for emergent reading and math skills (National Institute of Child Health and Human Development Early Child Care Research Network, 2000). The social-emotional learning (SEL) domain includes developmental skills—such as interacting positively with others and regulating emotions, attention and behavior—that promote on-task behavior and executive functioning (Rhoades, Warren, Domitrovich, & Greenberg, 2011). Children without the requisite SEL skills and who display behavior problems are less likely to engage in appropriate learning activities or exhibit a desire to learn and succeed (Wentzel, 1993), and their teachers are less likely to provide them with instruction and positive feedback (McEvoy & Welker, 2000). Over time, as children act out when presented with difficult tasks, their teachers tend to withdraw demands and provide fewer learning opportunities (Arnold et al., 1999; Carr, Taylor, & Robinson, 1991). In one of the few longitudinal studies of young Latino children's academic development, children with problem behavior in preschool had lower academic achievement in first grade (Oades-Sese, Esquivel, Kaliski, & Maniatis, 2011).

Overall, though, Latino students appear to be at relatively low risk for early behavior problems, in part because of a strong cultural emphasis on cultivating interpersonal (e.g., social) skills beginning at a very young age (Galindo & Fuller, 2010; Guerrero et al., 2013). In contrast, in the earliest years of schooling, Latinos lag behind their non-Latino peers across various indicators of academic school readiness. Latino pre-K students have lower levels of letter, number and shape recognition (Aud et al., 2010), and Latino kindergarten students are rated as less engaged and attentive (Llagas, 2003). Latino students score significantly below the national average on standardized achievement tests (Fryer & Levitt, 2004; Lee & Burkam, 2002). By fourth grade, 78% of Latino students fall below the proficient range on national standards of math and reading, and these rates remain stable through Grade 12, when 80% of Latinos have not reached proficiency in math and reading (Hemphill, Vanneman, & Rahman, 2011). These achievement gaps, which have remained relatively intractable over time (Hemphill et al., 2011), have dire implications for high school graduation rates (Alexander et al., 2001).

The Present Study

In response to the compelling need to identify malleable predictors that may be targeted in interventions to support the academic success and high school completion of Latino students, the present study uses a longitudinal design to examine early academic achievement and its relation to parenting practices, conceptualized as cultural socialization messages and the authoritative/authoritarian practices that reinforce them, that are shaped by mothers' acculturative status. First, we describe academic achievement at the end of first grade in a sample of MA and DA students in New York City. Second, we test a cultural framework of Latino parenting to identify predictors of student academic achievement at the end of first grade. In all analyses, we consider the role of ethnicity, gender, and grade.

Ethnicity

Considerable diversity characterizes the Latino population, which is made up of approximately 52 million immigrant and later-generation families from more than a dozen Spanish-speaking countries of origin. To provide a nuanced understanding of the pan-Latino population, the present study sampled purposively and exclusively from the Mexican- and Dominican-origin populations. The majority of U.S. Latinos (65%) come from Mexico, and though they have not historically resided in the Northeast, MAs are poised to quickly become the largest subgroup in New York City (NYC; Bergad, 2011). In contrast, the Dominican population has long represented one of the largest subgroups in NYC, where 1 in 5 Latinos is DA. The two groups differ significantly aeross various contextual characteristics, and citywide statistics show lower academic achievement in MA relative to DA students (Sáenz & Ponjuan, 2011).

Gender

Second, we examine gender differences, given the well-established disparities in achievement that favor girls nationally and within ethnic groups (NCES, 2000). Compared to girls, boys are more likely to repeat a grade, be suspended or expelled, be referred for special education, and be diagnosed with behavior problems that interfere with learning. The cumulative toll of these obstacles contributes to lower rates of high school graduation and postsecondary education among Latino males than females (Sáenz & Ponjuan, 2011). For example, only 57% of Latino boys who attend NYC public schools graduate from high school, compared with 67% of Latina girls (Villavicencio, Bhattacharya, & Guidry, 2013).

Grade

Finally, we examine differences between students who entered their assigned elementary school in prekindergarten versus kindergarten. The benefits of pre-K have been described extensively in the literature (Gormley, Gayer, Phillips, & Dawson, 2005), but only about half of Latino children attend pre-K (Espinosa, 2008). Estimates are even lower for children from Spanish-speaking Latino families (Espinosa, 2008), making the grade in which students start formal schooling especially pertinent in immigrant samples.

Based on these literatures, we explored mean-level differences based on ethnicity, gender, and grade in our first aim of describing academic achievement at the end of first grade. We expected DA students, girls, and children who entered school as pre-K students to be higher achieving than MA students, boys, and children who entered school as kindergarten students. For Aim 2 (model testing), we tested a 4-step sequential model which considers maternal acculturative status as a predictor of children's academic achievement, through its association with parenting (cultural socialization messages, parenting practices) and, in turn, the association of parenting with student school readiness, separately for MA and DA.

Method

Participants

Data were drawn from a longitudinal study of the early development of Latino children conducted by the authors in NYC. The study took place in 24 public elementary schools, where eligible 4- to 5-year-old children were enrolled between 2010–2013. Children were eligible if their mothers identified as Mexican (MA) or Dominican (DA) and if they were newly enrolled in prekindergarten or kindergarten in one of our partner schools. The final sample included 414 MA and 336 DA students and their mothers (N = 750 mother-child pairs). Consistent with previous reports on these two subsamples, MA (n = 414) and DA (n = 336) families differed on several demographic characteristics (see Table 1). Specifically, MA mothers were more likely to be immigrant (p < .001), have less than a high school education (p < .001), and live in poverty (p < .001) in comparison with DA mothers. MA children were younger than DA children (p < .001), and there were no significant difference between the MA and DA samples on child gender.

Measures

Demographic form. A comprehensive demographic form was administered to mothers. Child age and mother's age was measured in years as a continuous variable at baseline. Child gender was coded as boy (= 0) or girl (= 1). Mothers' self-identified ethnicity was coded as Mexican American (= 0) or Dominican American (= 1). Immigrant status for mothers and children was coded as U.S.-born (= 0) or immigrant (= 1). Children's language preference was coded as English (= 0) or Spanish (= 1), depending on the child's choice of language for the assessment at baseline. Mother's English and Spanish proficiency was measured using the language competence subscales of the AMAS (described further below; Zea, Asner-Self, Birman, & Buki, 2003). Nine items for each language that assessed expressive and receptive, as well as written and oral, language competence across contexts (e.g., work, home) were rated from 1 (not at all) to 4 (extremely well). The items were averaged for each language to create an English and a Spanish language competence scale. Mother's education level was coded as high school graduate or less (= 0) or college or more (= 1). Mother's employment was coded as not working for pay (= 0) or working for pay (= 1). To calculate a family income-to-needs ratio, we considered income relative to number of family members living in the home for whom the mother was financially responsible or with whom she was sharing household expenses using the federal poverty guidelines.

Acculturative status. The Abbreviated Multidimensional Acculturation Scale (AMAS; Zea et al., 2003) was used as a self-report measure of acculturative status (i.e., acculturation, enculturation). The AMAS can be used with any ethnic group; the use of the term *status* recognizes that acculturation/enculturation change over time and that the current measure reflects mothers' acculturation/enculturation at a specific point in time. This study used two domains of the AMAS: identity and cultural knowledge, both of which are measured for the culture of origin (enculturation) as well as mainstream/"U.S. American" culture (acculturation), resulting in four subscales (2 of enculturation and 2 of accultura-

Table 1
Descriptive Characteristics of the Sample, by Child Ethnicity, Gender, and Grade

	Descriptive statistics[a] *M* ± *SD*/*n* (%)	Ethnicity differences *M* ± *SD*/*n* (%)		Gender differences *M* ± *SD*/*n* (%)		Grade differences *M* ± *SD*/*n* (%)	
Characteristic		MA[b]	DA[c]	Boys[d]	Girls[e]	Pre-K[f]	K[g]
Child age	4.43 ± .60	4.39 ± .60	4.48 ± .59	4.46 ± .60	4.39 ± .60	3.88 ± .33	4.85 ± .38
		(*t* test = -2.05^{*})		(*t* test = 1.52)		(*t* test = -36.74^{***})	
Child gender (boys)	367 (48.9)	198 (47.8)	169 (50.3)	—	—	151 (46.3)	216 (50.9)
		(χ^2 = .45)		—		(χ^2 = 1.58)	
Child foreign-born	55 (7.5)	8 (2.0)	47 (14.4)	32 (9.0)	23 (6.1)	15 (4.7)	40 (9.7)
		(χ^2 = 39.64^{***})		(χ^2 = 2.14)		(χ^2 = 6.26^{**})	
Mothers' age	32.23 ± 6.73	31.12 ± 5.66	33.61 ± 7.64	32.54 ± 7.19	31.94 ± 6.26	31.92 ± 6.01	32.48 ± 7.24
		(*t* test = -4.94^{***})		(*t* test = 1.20)		(*t* test = −1.11)	
Child language preference (English)	375 (50.5)	145 (35.3)	230 (69.3)	185 (51.1)	190 (49.9)	137 (42.9)	238 (56.1)
		(χ^2 = 84.91^{***})		(χ^2 = .11)		(χ^2 = 12.66^{***})	
Mothers' ethnicity (MA)	414 (55.2)	—	—	198 (54.0)	216 (56.4)	185 (56.7)	229 (54.0)
		—		(χ^2 = .45)		(χ^2 = .56)	
Mother education (% ≤ high school)	541 (72.7)	378 (92.0)	163 (49.0)	267 (73.4)	274 (72.1)	218 (66.9)	323 (77.3)
		(χ^2 = 171.61^{***})		(χ^2 = .15)		(χ^2 = 9.99^{**})	
Mother works for pay	345 (46.7)	126 (31.0)	219 (66.0)	170 (47.0)	175 (46.4)	128 (39.5)	217 (52.3)
		(χ^2 = 90.02^{***})		(χ^2 = .88)		(χ^2 = 11.95^{**})	
Mother foreign-born	686 (92.2)	405 (98.5)	281 (84.4)	338 (93.1)	348 (91.3)	299 (92.0)	387 (92.4)
		(χ^2 = 51.28^{***})		(χ^2 = .81)		(χ^2 = .03)	
Mothers' English proficiency	2.15 ± .86	1.80 ± .59	2.58 ± .94	2.12 ± .84	2.18 ± .88	2.22 ± .88	2.10 ± .84
		(*t* test =-13.28^{***})		(*t* test =-81)		(*t* test = 1.97^{*})	
Mothers' Spanish proficiency	3.63 ± .51	3.58 ± .55	3.70 ± .45	3.64 ± .51	3.63 ± .51	3.64 ± .48	3.63 ± .53
		(*t* test =-3.32^{**})		(*t* test = .25)		(*t* test = .37)	
Family poverty status	498 (70.2)	317 (82.1)	181 (56.0)	247 (71.4)	251 (69.1)	215 (70.3)	283 (70.2)
		(χ^2 = 57.25^{***})		(χ^2 = .43)		(χ^2 = .00)	

Note. Dash means that no bivariate analysis occurred. MA = Mexican American; DA = Dominican American; Pre-K = Pre-kindergarten; K = Kindergarten. Because a full information procedure (FIML) estimates parameters for missing data based on all available data instead of imputing, the number of cases differs.

[a] n = 709 to 750. [b] n = 386 to 414. [c] n = 323 to 336. [d] n = 346 to 367. [e] n = 363 to 383. [f] n = 306 to 326. [g] n = 403 to 424.

[*] p < .05. [**] p < .01. [***] p < .001.

tion). The 24 items are rated from "not at all" (= 1) to "extremely well" (= 4). Sample items include "Being U.S. American/Dominican/Mexican plays an important part in my life," (identity); "How well do you know the national heroes of U.S./Mexico/the Dominican Republic?" (cultural knowledge). Raw scores of six items were used to specify each latent variable of subscale (i.e., U.S. identity, ethnic identity, U.S. cultural knowledge, and ethnic cultural knowledge). Internal consistencies were high for all subscales for both groups (MA: α range .89–.91; DA: α range .91–.92).

Cultural socialization messages. To assess socialization messages, mothers completed the Cultural Socialization of Latino Children (CSLC; Calzada et al., 2012), a measure that taps into the behavioral manifestations of the cultural value of *respeto* and the U.S. American value of independence (Calzada et al., 2010). Two factor-derived scales drawn from a principal component factor analysis with Varimax rotation were used in the present study: the *respeto* scale included 5 items (e.g., "I tell my child to show respect by greeting elders politely."), and the independence scale included 6 items (e.g., "I teach my child to share his own ideas and opinions."). Raw scores of 5 items and 6 items were used to specify latent variables of *respeto* and independence, respectively. Items are rated on a Likert scale from "*strongly disagree* (= 1)" to "*strongly agree* (= 5)." Internal consistency was adequate for both scales and was similar for the MA and DA samples (α were .72 and .73 for MA, and.78 and .76 for DA, for *respeto* and independence, respectively).

Parenting practices. The Parenting Styles and Dimensions (PSD; Robinson, Mandleco, Olsen, & Hart, 1995) was administered as a self-report measure of parenting practices corresponding to Baumrind's (1995) authoritarian and authoritative parenting style constructs. Parents respond to each item on a 5-point Likert scale anchored by "*never* (= 1)," "*once in a while* (= 2)," "*about half the time* (= 3)," "*very often* (= 4)," and "*always* (= 5)." The PSD has been shown to have strong face validity, concurrent validity, and predictive validity for schoolchildren, including preschool students, in the U.S. (Olivari, Tagliabue, & Confalonieri, 2013). The PSD has been standardized for parents of young children, and has been used with samples of various ethnic backgrounds including Latina mothers from Puerto Rico, the Dominican Republic and Mexico (Calzada & Eyberg, 2002; Calzada et al., 2012; Calzada et al., 2015). This study used the authoritative and authoritarian parenting scales. The authoritative parenting scale consists of three dimensions: connection (5 items; e.g., "I am responsive to my child's feelings or needs"), regulation (5 items: e.g., "I explain to my child how I feel about his or her good and bad behavior"), and autonomy granting (5 items; e.g., "I take my child's desires into account before asking him or her to do something"). The authoritarian scale also consists of three dimensions: physical coercion (4 items; e.g., "I use physical punishment as a way of disciplining my child"), verbal hostility (4 items; e.g., "I yell or shout when my child misbehaves"), and nonreasoning/punitive (4 items; e.g., "I punish by taking privileges away from my child with little if any explanation"). Average scores were calculated for these six dimensions, which were used as the specified latent variables for authoritative and authoritarian parenting. Alpha coefficients were acceptable to good (authoritative: MA = .86; DA = .83; authoritarian: MA = .69; DA = .66).

School readiness.

Academic school readiness. The Developmental Indicators for the Assessment of Learning-Third edition (DIAL-3; Mardell-Czudnowski & Goldenber, 1998) is an individually administered test that was administered as a measure of academic school readiness. The DIAL-3, and its abbreviated version, the Speed DIAL (used in the present study), assess motor, conceptual, and language development that is considered the foundation for successful academic learning. Ten concepts are tested including copying (motor), identifying colors (concepts), and letter naming (language). The DIAL has well-established psychometric properties, includes indicators of potential developmental delays and is available in Spanish and English. The Speed DIAL, which was used in the present study, consists of three subscales (motor, concepts, and language subscales) and was designed to be used with children ages 3.0-6.11. The Speed DIAL is predictive of math and English test scores in third grade (Walk, 2005). In the present study, average scores were calculated for each motor, concepts, and language subscales to specify a latent variable of academic readiness. On the Speed DIAL, the means in the standardization sample were: 20.0 (6.6) for children ages 4–0 through 4–5; 23.8 (6.0) for children ages 4–6 through 5–0; 27.2 (6.9) for children ages 5–0 through 5–5; and 30.9 (6.0) for children ages 5–6 through 6.

SEL school readiness. The Behavior Assessment System for Children-2 (BASC-2; Reynolds & Kamphaus, 2004), a measure of social-emotional and behavioral functioning with well-established psychometric properties, was administered to teachers. Teachers rate 139 items in terms of how often the child has engaged in a behavior during the past 4 weeks on a 4-point scale (*never* = 0; *sometimes* = 1; *often* = 2; and *almost always* = 3), and *T* scores are calculated based on child age (M = 50, SD = 10). In the present study, two latent variables were used: *teacher-rated adaptive behaviors* based on three measured variables (i.e., social skills, adaptability, communication) and *teacher-rated total problem behaviors* based on three measured variables (i.e., the depression, aggression, hyperactivity subscales of the BASC).

Child academic achievement. The Kaufman Test of Educational Achievement, Second Edition, Brief Form (KTEA-II; Kaufman & Kaufman, 2005) provides a quick, reliable estimation of global academic skills by assessing the achievement domains of reading, mathematics, and written language. Standardization of the KTEA included a Latino subsample. The Reading subtest includes word recognition (27 items) and reading comprehension (46 items). The Math and Writing subtests consist of 67 and 46 items, respectively. The KTEA-II Brief Form also derives a Brief Achievement Composite score. Standard scores are based on a mean of 100 and a standard deviation of 15. In the present study, children who scored >1 *SD* below the mean were categorized at risk. For model testing, we created a latent variable of child academic achievement using three measured variables of mean scores corresponding to the reading, math, and writing subtests of the KTEA.

Procedure

Enrollment of school partners. All data were drawn from schools that served MA and DA students in NYC. Public elementary schools were approached for partnership in the project via informational letters and phone calls if they (a) housed a universal pre-K program with at least 2 pre-K classes, and (b) had at least 20% Latino students according to Department of Education statistics. The 24 partner schools were classified as either a "Mexican" (n = 13) or a "Dominican" (n = 11) school depending on the predominant ethnic group of its Latino students as determined by neighborhood-level census data because Department of Education records do not have information on students' country of origin; all MA participants were drawn exclusively from "Mexican" schools and all DA participants were drawn exclusively from "Dominican" schools. Importantly, MA and DA schools did not differ on the contextual characteristics measured in the present study (Calzada et al., 2015). Both MA and DA students attended large public elementary schools (>650 students per school). The schools were highly segregated, with an average of 95% students of color and 95% eligible for free or reduced lunch.

Participant enrollment. We sampled from students entering their zoned public elementary school, whether as pre-K or kindergarten students, to characterize Latino children's experiences during the transition to formal schooling. Participant enrollment took place exclusively in the initial 3-month period of the school year. At partner schools, research staff, fluent in Spanish and English, attended parent meetings and were present during daily school drop-off and pick-up times to inform mothers of the study. The recruitment rate was 79% (range across schools: 59 – 98%), with no differences between MA and DA schools.

Data collection. As part of a larger, longitudinal study, mothers were interviewed, teachers completed questionnaires, and children were assessed in the fall when they entered school (i.e., in pre-K or kindergarten) and in the spring as they completed first grade. In the present study, data on predictor variables (mothers' acculturative status, parenting, student school readiness) were collected at baseline (i.e., in the Fall when children first entered school) and data on outcome variables (academic achievement) were collected at the end of first grade.

Participant mothers were scheduled for an appointment with a bilingual research assistant at their child's school. Mothers were asked which language they preferred to be interviewed in and were consented before beginning any research activities; all forms and measures were available in both Spanish and English. The majority of mothers (98% of MA and 76% of DA) chose to be interviewed in Spanish. Interviews lasted approximately 2 hours and mothers received a stipend for their participation. All child testing occurred during the school day at a time agreeable to the teacher. For the baseline assessment, testing was conducted by a bilingual research assistant in an hour-long session (broken into shorter sessions when needed) in the child's primary language (determined based on mother report, child report and research staff observation of child's language use in informal interactions). All children were at least 4 years old at the time of testing. Approximately 50% of MA and DA children (both MA and DA) were tested in Spanish. The final assessment in the spring of first grade followed the same procedure, except that all children were tested in English, the language in which they were evaluated by schools.

After the mother and child assessments, teachers of participating children were contacted (all mothers consented to the collection of teacher report) and their consent was obtained in person. Consenting teachers (92%) completed a teacher report packet on the child (e.g., child behavior) and mother (e.g., parent involvement) for each participating child in that classroom. Most (93%) consenting teachers returned their packets, resulting in 202 teachers who provided data on 368 pre-K and kindergarten classrooms and 702 (93.6%) participant children. There were no differences on demographics (i.e., ethnicity, mother education, marital status, family poverty, child gender), or parent-rated child functioning (i.e., externalizing, internalizing problems) between children who had teacher ratings and those who did not (p > .05).

Analytic Approach

We used SPSS and AMOS software programs for the analyses in the present study. For our first aim, to describe academic achievement and its predictors, we conducted descriptive and bivariate statistics on the variables for the full sample and for subgroups based on ethnicity, gender, and grade when children entered school (i.e., in pre-K or kindergarten).

For our second aim, to test a model of parenting and academic achievement, we used two time points: baseline data for predictor variables (acculturative status, socialization messages, parenting practices, school readiness) and follow-up data collected at the end of the child's first grade school year for our outcome variable (academic achievement test scores). To account for missingness, we used a full information procedure (FIML). Instead of imputing for missing data, FIML derives likelihood functions based on all available data and then estimates parameters for all data points (Enders & Bandalos, 2001).

Next, we examined bivariate correlations between the study variables for MA and DA students separately (see Table 4). We found significant associations consistent with our conceptual model, and although there were many ethnic group similarities in the pattern of associations, there were also some differences. Thus, we conducted a measurement invariance test between MAs and DAs on all measures.

We then tested a structural equation model (SEM; see Figure 1) using the maximum likelihood estimation method. Although individual items were used to specify the latent variables of acculturative status and cultural socialization messages, subscale scores from each of the measures were used to specify the latent variables of parenting practices, school readiness, and academic achievement in the model. We chose an SEM analytic approach over multiple regression analyses because of its ability (a) to examine sequential links from mothers' acculturation and enculturation, to socialization messages, to parenting practices, to school readiness, and finally to academic achievement and (b) to correct for measurement error and residual error (Byrne, 2001). Our data consist of 750 children within 388 classrooms and 24 schools. We did not use multilevel analysis for two reasons. First, the intraclass correlation (ICC) for academic achievement at the school level was .058 (τ_{00} = 71.89; p > .05) in MA and .045 (τ_{00} = 77.34; p > .05 in DA. According to Lee (2000), multilevel modeling is recommended when ICC is greater than .10. Also, the '30/30 rule,' which stipulates that a sample of at least 30 groups with at least 30 individuals per group is needed for multilevel modeling (Kreft, 1996), was not met in the present study. More than half of the classrooms (195 classes) had one participating child, and another quarter (97 classes) had two participating children.

To test our model, we used multigroup structural equation modeling and controlled for child's grade when they entered school, child gender, child's language of assessment at baseline, and family poverty (i.e., income-to-needs ratio). First, our conceptual model was tested with MA and DA children separately to establish the baseline model fit and examine possible structural differences between the two groups. Four indices were used to assess the goodness of fit of the hypothesized model to the actual data in multi-Group SEM per Keith (2014) and Kline (2005): chi-square ($\chi^2 > .05$ or χ^2/df ratio less than 3.0), Adjusted Root Mean Square Error of Approximation (adjusted RMSEA $< .05$), Comparative Fit Index (CFI $> .90$), and Tucker-Lewis index (TLI $> .90$). Then, we conducted a measurement invariance test, a prerequisite of multigroup structural equation modeling, between ethnic groups (i.e., MAs and DAs). The measurement invariance test set factor loadings from all latent to measured variables to be the same across groups. The test determines whether measures are functioning in a comparable way across groups. We used the chi-square difference test, RMSEA, and TLI to compare the fit of nested models (Hong & Ho, 2005; Hong, Malik, & Lee, 2003). Once measurement invariance was supported, we then conducted multigroup structural equation modeling by ethnicity to test where structural path differences occurred. All the constraints from the measurement invariance test were retained, and structural paths were constrained one at a time. A constraint of structural path means that that path will be freely estimated but that the unstandardized path will be constrained to be equal across groups (Keith, 2014). A statistical significant degradation in the fit of the model indicates a significantly different path coefficient across groups.

Results

Preliminary Analyses

Table 1 presents the baseline characteristics for the full sample and for the three subgroups based on ethnicity (55.2% MA), gender (48.9% boys), and grade (i.e., when they entered school, in pre-K or kindergarten; 56.5% kindergarten). Compared with DA parents, MA parents were younger (31.12 vs. 33.61 years) and more likely to be foreign-born (98.5% vs. 84.4%), less educated (92.0% vs. 49.0% high school graduates), and living in poverty (82.1% vs. 56.0%). There were no differences based on child gender. However, baseline characteristics showed that compared with those in pre-K, children in kindergarten were more likely to be foreign-born (9.7% vs. 4.7%) and their mothers were more likely to have low educational levels (77.3% vs. 66.9% high school graduates) and to work outside of the home (52.3% vs. 39.5%).

Table 2 shows descriptive statistics for parenting variables as well as bivariate statistics with the endogenous variable (academic achievement in first grade). Mothers reported high levels of ethnic identity and ethnic cultural knowledge, and moderate levels of U.S. American identity and cultural knowledge; high levels of socialization to both *respeto* and independence; and high levels of authoritative and low levels of authoritarian parenting. Compared to MA mothers, DA mothers reported significantly higher levels of U.S. identity (2.99 vs. 2.19), U.S. cultural knowledge (2.44 vs. 1.63), ethnic cultural knowledge (2.81 vs. 2.63), socialization messages of independence (4.49 vs. 4.38), and authoritative parenting (4.51 vs. 4.25). There were no significant differences in socialization messages of independence or parenting practices (authoritative, authoritarian) between mothers of boys and girls, or mothers of pre-K and kindergarten children.

There were robust ethnic group differences on school readiness (see Table 3). Specifically, DA children were rated as engaging in more problem behaviors (48.54 vs. 46.13) but also in more adaptive behaviors (47.54 vs. 45.57) in the classroom, and they scored higher on a test of academic readiness (23.86 vs. 20.35) than MA children. In comparing boys and girls, boys showed more problem behaviors (48.15 vs. 46.24) and less adaptive behaviors (45.58 vs. 47.28 points) in the classroom than girls. Finally, children in kindergarten scored higher in academic readiness than children in pre-K (age-adjusted scores of 25.83 vs. 16.75). Notably, there were no significant differences in mean-level scores of first grade academic achievement by child ethnicity or gender. However, at the end of first grade, children from the kindergarten cohort were more likely than those from the pre-K cohort to have lower mean-level scores of first grade achievement (96.28 vs. 98.50) and to have

Table 2
Descriptive Analysis of Latent Variables of the Sample, by Child Ethnicity, Gender, and Grade

Variable	Descriptive statistics[a] *M* ± *SD*	Ethnicity differences *M* ± *SD*			Gender differences *M* ± *SD*			Grade differences *M* ± *SD*		
		MA[b]	DA[c]	*t* test	Boys[d]	Girls[e]	*t* test	Pre-K[f]	K[g]	*t* test
U.S. identity	2.55 ± .93	2.19 ± .85	2.99 ± .84	−12.9***	2.61 ± .92	2.50 ± .95	1.6	2.52 ± .94	2.58 ± .93	−.8
Ethnic identity	3.87 ± .37	3.91 ± .31	3.82 ± .43	3.3**	3.89 ± .35	3.85 ± .39	1.3	3.85 ± .42	3.88 ± .32	−.9
U.S. cultural knowledge	1.99 ± .74	1.63 ± .53	2.44 ± .71	−17.29***	1.98 ± .74	2.00 ± .74	−.3	2.01 ± .74	1.97 ± .74	.7
Ethnic cultural knowledge	2.71 ± .77	2.63 ± .74	2.81 ± .79	−3.2**	2.73 ± .77	2.70 ± .77	.5	2.67 ± .76	2.75 ± .78	−1.3
Socialization of independence	4.43 ± .43	4.38 ± .42	4.49 ± .43	−3.2**	4.44 ± .43	4.42 ± .43	.9	4.43 ± .41	4.43 ± .44	−.1
Socialization of *respeto*	4.33 ± .52	4.30 ± .51	4.36 ± .52	−1.5	4.35 ± .51	4.30 ± .52	1.4	4.28 ± .52	4.36 ± .51	−2.2*
Authoritative practices	4.12 ± .64	4.25 ± .67	4.51 ± .53	−5.8***	4.36 ± .64	4.37 ± .61	−.2	4.37 ± .59	4.37 ± .65	.0
Authoritarian practices	1.75 ± .46	1.63 ± .52	1.49 ± .47	3.9***	1.60 ± .51	1.53 ± .48	1.9	1.57 ± .48	1.56 ± .51	.3

Note. MA = Mexican American; DA = Dominican American; Pre-K = Pre-kindergarten; K = Kindergarten; Because a full information procedure (FIML) estimates parameters for missing data based on all available data instead of imputing, the number of cases differs.
[a] $n = 733$ to 745. [b] $n = 401$ to 411. [c] $n = 328$ to 334. [d] $n = 356$ to 364. [e] $n = 372$ to 381. [f] $n = 317$ to 326. [g] $n = 414$ to 419.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 3

School Readiness and Academic Achievement by Child Ethnicity, Gender, and Grade

Variable	Full sample[a] *M* ± *SD* (%)	MA[b] *M* ± *SD* (%)	DA[c] *M* ± *SD* (%)	*t* test or χ^2	Boys[d] *M* ± *SD* (%)	Girls[e] *M* ± *SD* (%)	*t* test or χ^2	Pre-K[f] *M* ± *SD* (%)	K[g] *M* ± *SD* (%)	*t* test or χ^2
					SEL school readiness					
Problem behaviors	47.19 ± 7.37	46.13 ± 6.50	48.54 ± 8.17	−4.31***	48.15 ± 8.21	46.24 ± 6.31	3.43**	46.05 ± 6.22	48.08 ± 8.05	−3.61***
Adaptive behaviors	46.44 ± 7.95	45.57 ± 7.62	47.54 ± 8.24	−3.22**	45.58 ± 7.63	47.28 ± 8.18	−2.78**	45.93 ± 7.69	46.83 ± 8.14	−1.46
					Academic school readiness					
Speed DIAL-3	21.92 ± 7.37	20.35 ± 7.24	23.86 ± 7.07	−6.65***	21.71 ± 7.42	22.11 ± 7.32	−.74	16.75 ± 6.16	25.83 ± 5.58	−20.73***
					Academic achievement at the end of first grade (KTEA)					
KTEA	97.15 ± 12.43	96.73 ± 11.33	97.73 ± 13.81	−.90	96.99 ± 13.07	97.29 ± 11.85	−.28	98.50 ± 11.20	96.28 ± 13.10	2.05*
% at risk on KTEA	(18.5%)	(16.3%)	(21.7%)	(2.42)	(20.7%)	(16.7%)	(1.36)	(11.8%)	(22.9%)	(10.01**)

Note. DIAL-3 was measured by the abbreviated version of the Developmental Indicators for the Assessment of Learning-Third edition (Speed DIAL-3) at baseline. KTEA was measured by the Kaufman test of Educational Achievement, Second Edition, Brief Form (KTEA-II) at the end of first grade. MA = Mexican American; DA = Dominican American; Pre-K = Pre-kindergarten; K = Kindergarten. Because a full information procedure (FIML) estimates parameters for missing data based on all available data instead of imputing, the number of cases differs.

[a] *n* = 518 to 744. [b] *n* = 301 to 411. [c] *n* = 217 to 333. [d] *n* = 242 to 364. [e] *n* = 276 to 381. [f] *n* = 203 to 326. [g] *n* = 315 to 419.

* *p* < .05. ** *p* < .01. *** *p* < .001.

Table 4

Correlations Among Latent Variables by Ethnicity Group

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1. U.S. identity	1	−.073	.318***	−.164***	.075	.032	.086	−.027	.038	.067	−.018	−.085
2. Ethnic identity	−.077	1	−.210***	.293***	.007	.173**	.059	.011	−.043	.100	.024	−.051
3. U.S. cultural knowledge	.216***	−.021	1	−.071	.119*	−.083	.122*	.029	.151**	−.030	.105	.082
4. Ethnic cultural knowledge	−.066	.162**	.222***	1	.168**	.340***	.133*	−.119*	−.068	−.021	−.052	−.034
5. Socialization of independence	−.056	.081	.083	.187***	1	.532***	.398***	−.153**	−.157	.025	.033	−.021
6. Socialization of *respeto*	−.099*	.081	−.023	.158**	.546***	1	.295***	−.137*	−.066	.017	−.010	−.144*
7. Authoritative practices	−.062	.030	.129*	.263***	.410***	.302***	1	−.127*	−.042	.090	.056	.076
8. Authoritarian practices	.019	.063	−.106*	.060	−.026	−.024	−.020	1	.104	−.016	.039	.013
9. Problem behaviors	.013	.039	−.079	−.055*	−.035	.073	.032	.129*	1	−.415***	−.062	−.165*
10. Adaptive behaviors	−.034	−.027	.123*	.026	.037	−.091	.032	−.062	−.239***	1	.395***	.387***
11. Academic readiness	−.030	−.060	.068	.020	.071	−.028	.084	−.021	.055	.361***	1	.468***
12. Academic achievement at the end of first grade	−.085	−.047	.087	−.015	.010	−.006	.176**	−.004	−.163**	.371***	.361***	1

Note. Correlations for the Mexican sample presented below the diagonal; correlations for the Dominican sample presented above the diagonal.

* *p* < .05. ** *p* < .01. *** *p* < .001.

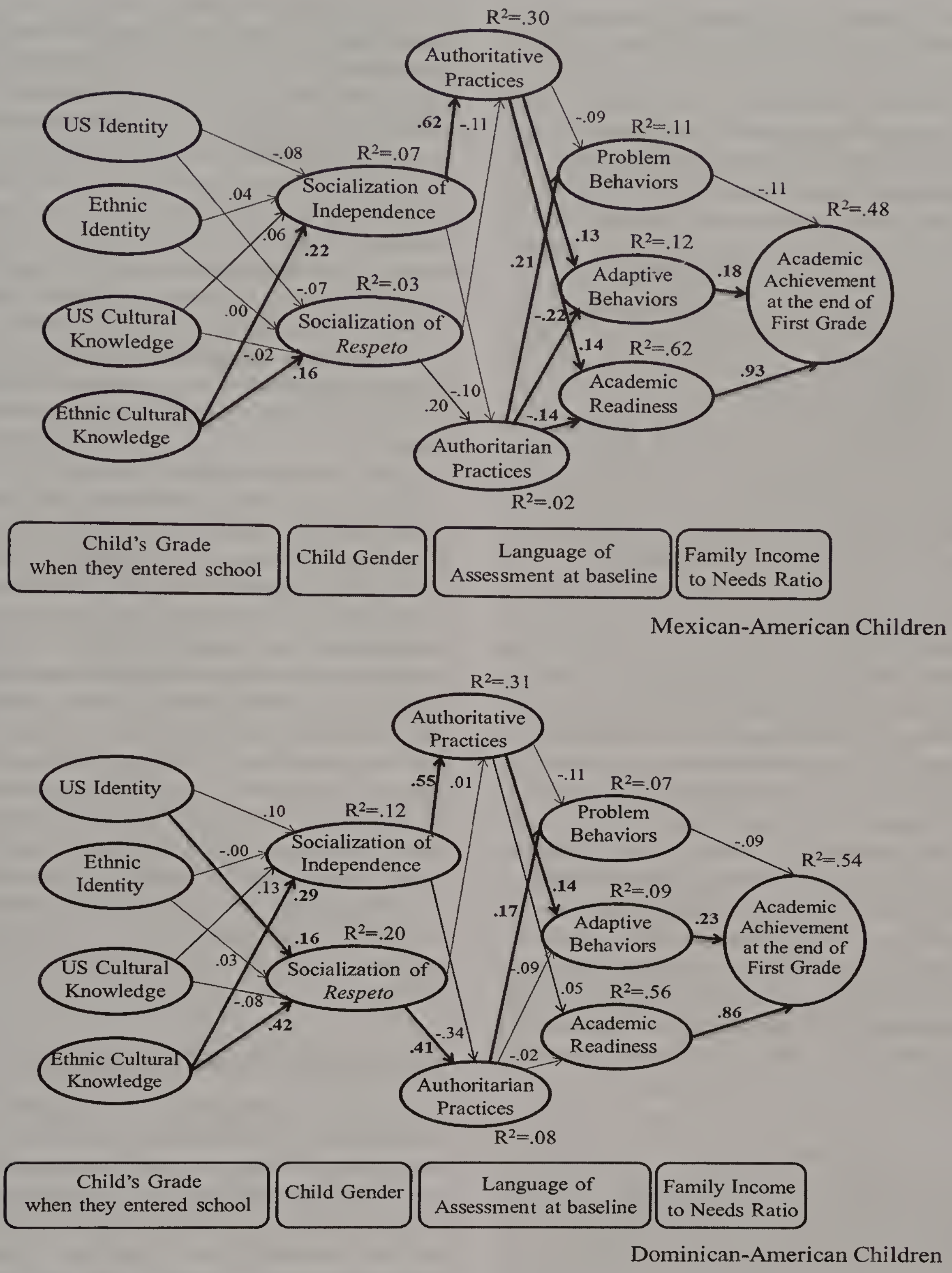


Figure 2. Structural model with standardized estimates. Bold lines and numbers indicate significant paths and standardized coefficients.

scores at least one standard deviation below the mean, or at academic risk (22.9% vs. 11.8%).

Multi-Group Structural Equation Modeling

The model was tested with MA and DA children separately. The fit indices of the structural model were acceptable (χ^2/df = 1.55, adjusted RMSEA = .04, TLI = .91, CFI = .92). The standardized coefficients and R^2 are presented in Figure 2. For MA students, mothers' ethnic cultural knowledge was positively related to their socialization of independence (standardized coefficient = .22) and *respeto* (standardized coefficient = .16); socialization of independence was associated with more authoritative parenting practices (standardized coefficient = .62); and more authoritative parenting was associated with better adaptive behaviors (standardized coefficient = .13) and higher academic readiness (standardized coefficient = .14) in students. Authoritarian parenting practices, in contrast, were associated with more problem behaviors (standardized coefficient = .21), fewer adaptive behaviors (standardized coefficient = −.22), and lower academic readiness (standardized coefficient = −.14) in students. Finally, academic readiness and adaptive behaviors at baseline predicted higher academic achievement at the end of first grade (standardized coefficient = .93 and .18, respectively).

For DA students, mothers' ethnic cultural knowledge was positively associated with their socialization of independence (standardized coefficient = .29) and *respeto* (standardized coefficient = .42), and their U.S. identity was positively associated with their socialization of *respeto* (standardized coefficient = .16). Socialization of independence was associated with more authoritative parenting practices (standardized coefficient = .55), whereas socialization of *respeto* was associated with more authoritarian parenting practices (standardized coefficient = .41). More authoritative parenting practices were associated with better adaptive behaviors (standardized coefficient = .14), and more authoritarian parenting practices were associated with more problem behaviors (standardized coefficient = .17) in students. Finally, academic readiness and adaptive behaviors were significant predictors of academic achievement (standardized coefficient = .86 and .23, respectively). The standardized direct and indirect effects on academic achievement are shown in the Appendix.

A measurement invariance test was conducted as a prerequisite of multigroup structural equation modeling. For a measurement invariance test between MAs and DAs, we constrained the factor loadings to be equal. The constraints increased the chi-square value from 4,543 to 4,613 and were statistically significant at alpha .05. Although the chi-square difference test is often used to compare the fit of nested models, the additional consideration of RMSEA and TLI is recommended (Hong & Ho, 2005; Hong et al., 2003). The baseline model and the model with metric invariance showed the same TLI values (baseline model = .91; metric invariance model = .91) and adjusted RMSEA (baseline model = .04; metric invariance model = .04), indicating that metric invariance was in fact supported. Given that the condition of measurement invariance was met, we conducted a multigroup structural equation modeling by ethnicity. Two of the 12 the specified path coefficients were significantly different across ethnic groups: from U.S. identity to *respeto*, $\Delta\chi^2 = 7.05$, $p < .01$; and from cultural knowledge to *respeto*, $\Delta\chi^2 = 7.11$, $p < .01$.

Discussion

The present study examined academic achievement among Latino students, focusing on variations by ethnicity, gender and grade when they entered school, and tested a model of parenting as a predictor of Latino students' early achievement. To our knowledge, this is the first study to examine longitudinal associations between parenting and standardized achievement test scores in a sample of young (first grade) Latino students. Our findings highlight ethnic subgroup differences and the role of parenting, which was found to predict academic achievement indirectly, via academic and social-emotional school readiness skills in pre-K or kindergarten.

Early Academic Achievement Among Latino Students

In assessing academic achievement at the end of first grade, we found that the Latino students in our sample were generally achieving at grade level (i.e., mean of 96.56 on the KTEA) and this held true for MA and DA students, boys and girls, and students who started formal schooling in pre-K versus kindergarten. Past studies have found that Latino students enter pre-K and kindergarten with low levels of academic skills (Fryer & Levitt, 2004; Lee & Burkam, 2002). The achievement gap has been shown to narrow, however, between pre-K, kindergarten and first grade (Reardon & Galindo, 2009). Similarly, in the present study, mean-level descriptive statistics show that even students with limited academic readiness skills (e.g., MA students) or low social-emotional school readiness (e.g., boys) at the start of school were, as a whole, achieving at grade level by the spring of first grade.

Still, 18.5% of students were in the at-risk range (defined as 1 + *SD* below the mean on standardized testing) at the end of first grade, with notable group differences observed. First, children who entered formal schooling in kindergarten rather than pre-K were twice as likely to be at risk for academic underachievement at the end of first grade (23% compared with 12%). This pattern is consistent with past studies showing the importance of preschool for academic success, especially among students from non-English speaking backgrounds (Howes et al., 2008; Wong, Cook, Barnett, & Jung, 2008). Second, according to post hoc descriptive analyses on ethnic and gender subgroups, 28% of DA boys were at risk for academic underachievement at the end of first grade, compared with approximately 15% of MA boys, MA girls and DA girls. Future research is needed to examine the academic trajectories of diverse Latino students, with attention to what risk factors may disproportionately affect DA boys. For example, characteristics of the classroom, including student-teacher relationships, will be important to explore in future work (Hughes & Kwok, 2007; Suárez-Orozco, Suárez-Orozco, & Qin-Hillard, 2005). Also, though the question of gender socialization and schooling in young Latino students remains largely unexplored, there is some modest evidence that gender influences academic motivation and supports among Latino students at older ages (e.g., Alfaro, Umaña-Taylor, & Bámaca, 2006; Yowell, 2000; Williams, Alvarez, & Hauck, 2002).

Parenting and Academic Achievement

Our test of a conceptual model of parenting is consistent with past studies (see Calzada et al., 2010; Calzada et al., 2012; Calzada et al., 2015) in highlighting how maternal acculturative status is associated with parenting—including cultural socialization messages and authoritative/authoritarian practices. Unique to the present study is parenting as a predictor of children's academic achievement through school readiness in academic and social-emotional domains. Findings illustrate the protective role of authoritative parenting practices for both MA and DA students. Specifically, authoritative parenting was associated with better social-emotional school readiness in DA and MA students, and with academic school readiness in MA students. For all students, school readiness was associated with better academic achievement later on. Authoritarian parenting practices, on the other hand, were associated with lower social-emotional school readiness for DA and MA students and with lower academic readiness in MA students, though its link to later underachievement was significant only among MA students.

It is not clear from our data why these ethnic group differences exist, but our findings mirror inconsistencies in the broader literature on authoritarian parenting practices in Latino families (Calzada et al., 2012; Calzada et al., 2015; Gorman-Smith, Tolan, Henry, & Florsheim, 2000; Knight, Virdin, & Roosa, 1994). Thus, it is likely that the impact of parenting varies across developmental outcomes, time points and contexts (i.e., MA and DA families differed on almost all contextual variables measured in the present study). How these seemingly complex and dynamic effects of risk and protection may play out is an important area for ongoing research.

The field must also carefully consider the ways in which it characterizes Latino parents, who are often described as harsh and controlling based on Westernized notions of parenting (Cardona et al., 2000; Guilamo-Ramos et al., 2007; Rodriguez & Olswang, 2003). It is notable that although we relied on a Westernized typology of parenting (e.g., authoritarian/authoritative parenting) in the present study, we found high levels of (self-reported) authoritative and low levels of authoritarian parenting, especially among DA mothers. These results suggest that harsh, controlling parenting may not be the norm for all Latino parents and that in fact, authoritative parenting may well be culturally congruent, at least during the preschool years.

Still, consistent with past research (Calzada et al., 2010; Calzada et al., 2012; Calzada et al., 2015), culture was found to play an important role in parenting, with mothers' acculturative status associated with socialization messages and parenting practices. For both MA and DA mothers, ethnic cultural knowledge was associated with more socialization to the U.S. American value of independence and to the cultural value of *respeto*, and socialization messages of independence, in turn, was associated with more authoritative parenting. For DA mothers only, socialization to the cultural value of *respeto* was associated with more authoritarian parenting. Theoretically, socialization messages of independence (e.g., "I encourage my child to tell me when s/he disagrees with me") align with the use of authoritative parenting practices (e.g., "I give my child reasons why rules should be obeyed"), and socialization messages of *respeto* (e.g., "I correct my child when he or she does not offer to help elders") align with the use of authoritarian parenting practices (e.g., "I use physical punishment as a way of disciplining my child"; Calzada et al., 2010). Empirically, higher levels of acculturation *and* enculturation appear to be associated indiscriminately with higher levels of socialization to U.S. *and* Latino values. Perhaps both constructs capture, at least in part, a mother's underlying approach to social interactions (e.g., actively engaging with others, whether it is members of her culture, members of mainstream society, or her child; Calzada et al., 2012). We are currently exploring other ways of studying acculturation (e.g., by examining acculturative stress rather than acculturation levels) that may better inform family processes.

Limitations and Conclusions

There are several notable limitations to the present study. First, although we used longitudinal data to examine children's academic achievement, we relied on cross-sectional data when examining associations between acculturative status, socialization messages, parenting practices, and children's school readiness. Thus, although our data support a conceptual model in which acculturative status shapes socialization messages and parenting practices, we were not able to establish temporal ordering and causality cannot be inferred. Second, as noted above, our measure of parenting practices was based on a Western typology (i.e., authoritarian, authoritative), and some scholars have questioned its applicability to Latino families (Domenech Rodríguez, Donovick, & Crowley, 2009). Indeed, the authoritarian scale used in this study had modest reliability. Furthermore, we relied on mothers' self-report, which is subject to rater bias. Thus, although the present study suggests that research on authoritarian/authoritative parenting in Latinos can be useful in understanding developmental outcomes, more scholarship on Latino parenting theory is needed to ensure methodological rigor and cultural sensitivity. Third, our study used a more global index of problem behaviors because conceptually, "social emotional school readiness" is a domain that encompasses all problem behaviors, whether externalizing or internalizing. Problem behaviors were not a significant predictor of later academic achievement in the present study. Given that there could be differences in how internalizing behaviors are associated with parenting and school readiness relative to externalizing behavior problems, future research should explore problem domain-specific associations with school readiness and academic achievement.

Fourth, our model was limited to select family level variables, and did not consider other ecological variables that undoubtedly influence academic achievement, including characteristics of the classroom and school. Our model explained a modest amount of the variance in teacher-reported problem and adaptive behavior, indicating the presence of other variables in shaping child functioning in the classroom. For example, quality of instruction (i.e., teaching practices), cultural competence of school staff, bilingual education resources, and the demographic characteristics of schools are likely important determinants of school readiness and academic achievement (Caldas & Bankston, 1997; Ma & Klinger, 2000; Rivkin, Hanushek, & Kain, 2005). Outside of the classroom, the neighborhood context (e.g., poverty, safety, segregation) has also been found to influence academic achievement (Hanson et al., 2011; Lapointe, Ford, & Zumbo, 2007; Leventhal & Brooks-Gunn, 2000) and should be considered in future research with Latinos.

Finally, our sample was limited to MA and DA families, and study findings cannot be generalized to other Latino families. The tremendous diversity of the Latino population has been well-described in the literature, and although we accounted for some sources of heterogeneity (i.e., ethnicity, child gender, grade when they entered school), we were not able to account for others such as immigrant status (the vast majority of students had foreign-born parents) of students, important next steps in this line of research.

Although more work is needed to understand early academic achievement in Latino students, the present study makes a significant contribution to identifying malleable factors that may be associated with early achievement. Among low-income families, parenting practices represent a particularly important malleable factor because they have the potential to buffer children from the negative impact of socioeconomic challenges (Hill, 2001). Findings from the present study suggest that parenting that aims to encourage age-appropriate independence in young children and that relies on authoritative practices may promote the development of children's early academic and social-emotional competencies, with positive effects on academic achievement in later years.

References

Alexander, K., Entwisle, D., & Kabbani, N. (2001). The dropout process in life course perspective: Early risk factors at home and school. *Teachers College Record, 103,* 760–822. http://dx.doi.org/10.1111/0161-4681.00134

Alfaro, E. C., Umaña-Taylor, A. J., & Bámaca, M. Y. (2006). The influence of academic support on Latino adolescents' academic motivation. *Family Relations, 55,* 279–291. http://dx.doi.org/10.1111/j.1741-3729.2006.00402.x

Arnold, D. H., Ortiz, C., Curry, J. C., Stowe, R. M., Goldstein, N. E., Fisher, P. H., . . . Yershova, K. (1999). Promoting academic success and preventing disruptive behavior disorders through community partnership. *Journal of Community Psychology, 27,* 589–598. http://dx.doi.org/10.1002/(SICI)1520-6629(199909)27:5<589::AID-JCOP6>3.0.CO;2-Y

Aud, S., Fox, M. A., & KewalRamani, A. (2010). Status and trends in the education of racial and ethnic groups. NCES 2010–015. *National Center for Education Statistics.*

Baumrind, D. (1995). Child rearing dimensions relevant to child maltreatment. In R. Lerner (Ed.), *Child maltreatment and optimal care giving in social contexts* (pp. 55–73). New York, NY: Garland.

Bergad, L. W. (2011). *Mexicans in New York City, 1990–2009: A visual database.* Center for Latin American, Caribbean & Latino Studies, CUNY Graduate Center.

Brooks-Gunn, J., Guo, G., & Furstenberg, F. F., Jr. (1993). Who drops out of and who continues beyond high school? A 20-year follow-up of black urban youth. *Journal of Research on Adolescence, 3,* 271–294. http://dx.doi.org/10.1207/s15327795jra0303_4

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications and programming.* Mahwah, NJ: Erlbaum.

Cairns, R. B., & Cairns, B. D. (1994). *Lifelines and risks: Pathways of youth in our time.* Cambridge, UK: Cambridge University Press.

Caldas, S. J., & Bankston, C. (1997). Effect of school population socioeconomic status on individual academic achievement. *The Journal of Educational Research, 90,* 269–277. http://dx.doi.org/10.1080/00220671.1997.10544583

Calzada, E. J. (2010). Bringing culture into parent training with Latinos. *Cognitive and Behavioral Practice, 17,* 167–175. http://dx.doi.org/10.1016/j.cbpra.2010.01.003

Calzada, E., Barajas-Gonzalez, R. G., Huang, K. Y., & Brotman, L. (2015). Early childhood internalizing problems in Mexican-and Dominican-origin children: The role of cultural socialization and parenting practices. *Journal of Clinical Child and Adolescent Psychology.* Advance online publication. http://dx.doi.org/10.1080/15374416.2015.1041593

Calzada, E. J., & Eyberg, S. M. (2002). Self-reported parenting practices in Dominican and Puerto Rican mothers of young children. *Journal of Clinical Child and Adolescent Psychology, 31,* 354–363. http://dx.doi.org/10.1207/S15374424JCCP3103_07

Calzada, E. J., Fernandez, Y., & Cortes, D. E. (2010). Incorporating the cultural value of respeto into a framework of Latino parenting. *Cultural Diversity and Ethnic Minority Psychology, 16,* 77–86. http://dx.doi.org/10.1037/a0016071

Calzada, E. J., Huang, K. Y., Anicama, C., Fernandez, Y., & Brotman, L. M. (2012). Test of a cultural framework of parenting with Latino families of young children. *Cultural Diversity and Ethnic Minority Psychology, 18,* 285–296. http://dx.doi.org/10.1037/a0028694

Cardona, P. G., Nicholson, B. C., & Fox, R. A. (2000). Parenting among Hispanic and Anglo-American mothers with young children. *The Journal of Social Psychology, 140,* 357–365. http://dx.doi.org/10.1080/00224540009600476

Carr, E. G., Taylor, J. C., & Robinson, S. (1991). The effects of severe behavior problems in children on the teaching behavior of adults. *Journal of Applied Behavior Analysis, 24,* 523–535. http://dx.doi.org/10.1901/jaba.1991.24-523

Domenech Rodríguez, M. M., Donovick, M. R., & Crowley, S. L. (2009). Parenting styles in a cultural context: Observations of "protective parenting" in first-generation Latinos. *Family Process, 48,* 195–210. http://dx.doi.org/10.1111/j.1545-5300.2009.01277.x

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43,* 1428–1446. http://dx.doi.org/10.1037/0012-1649.43.6.1428

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8,* 430–457. http://dx.doi.org/10.1207/S15328007SEM0803_5

Ensminger, M. E., & Slusarcick, A. L. (1992). Paths to high school graduation or dropout: A longitudinal study of a first-grade cohort. *Sociology of Education, 65,* 95–113. http://dx.doi.org/10.2307/2112677

Espinosa, L. M. (2008). *Challenging common myths about young English language learners.* New York, NY: Foundation for Child Development.

Finn, J. D., Gerber, S. B., & Boyd-Zaharias, J. (2005). Small classes in the early grades, academic achievement, and graduating from high school. *Journal of Educational Psychology, 97,* 214–223. http://dx.doi.org/10.1037/0022-0663.97.2.214

Fryer, R. G., Jr., & Levitt, S. D. (2004). Understanding the black-white test score gap in the first two years of school. *The Review of Economics and Statistics, 86,* 447–464. http://dx.doi.org/10.1162/003465304323031049

Galindo, C., & Fuller, B. (2010). The social competence of Latino kindergartners and growth in mathematical understanding. *Developmental Psychology, 46,* 579–592. http://dx.doi.org/10.1037/a0017821

Garnier, H. E., Stein, J. A., & Jacobs, J. K. (1997). The process of dropping out of high school: A 19-year perspective. *American Educational Research Journal, 34,* 395–419. http://dx.doi.org/10.3102/00028312034002395

Gonzales, N. A., Germán, M., Kim, S. Y., George, P., Fabrett, F. C., Millsap, R., & Dumka, L. E. (2008). Mexican American adolescents' cultural orientation, externalizing behavior and academic engagement: The role of traditional cultural values. *American Journal of Community Psychology, 41,* 151–164. http://dx.doi.org/10.1007/s10464-007-9152-x

Gorman-Smith, D., Tolan, P. H., Henry, D. B., & Florsheim, P. (2000). Patterns of family functioning and adolescent outcomes among urban African American and Mexican American families. *Journal of Family Psychology, 14,* 436–457. http://dx.doi.org/10.1037/0893-3200.14.3.436

Gormley, W. T., Jr., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology, 41,* 872–884. http://dx.doi.org/10.1037/0012-1649.41.6.872

Guarnaccia, P. J., & Rodriguez, O. (1996). Concepts of culture and their role in the development of culturally competent mental health services. *Hispanic Journal of Behavioral Sciences, 18,* 419–443. http://dx.doi.org/10.1177/07399863960184001

Guerrero, A. D., Fuller, B., Chu, L., Kim, A., Franke, T., Bridges, M., & Kuo, A. (2013). Early growth of Mexican-American children: Lagging in preliteracy skills but not social development. *Maternal and Child Health Journal, 17,* 1701–1711. http://dx.doi.org/10.1007/s10995-012-1184-7

Guilamo-Ramos, V., Dittus, P., Jaccard, J., Johansson, M., Bouris, A., & Acosta, N. (2007). Parenting practices among Dominican and Puerto Rican mothers. *Social Work, 52,* 17–30. http://dx.doi.org/10.1093/sw/52.1.17

Hanson, M. J., Miller, A. D., Diamond, K., Odom, S., Lieber, J., Butera, G., . . . Fleming, K. (2011). Neighborhood community risk influences on preschool children's development and school readiness. *Infants & Young Children, 24,* 87–100. http://dx.doi.org/10.1097/IYC.0b013e3182008dd0

Harwood, R. L. (1992). The influence of culturally derived values on Anglo and Puerto Rican mothers' perceptions of attachment behavior. *Child Development, 63,* 822–839. http://dx.doi.org/10.2307/1131236

Haveman, R., & Wolfe, B. (1994). *Succeeding generations: On the effects of investments in children.* New York, NY: Russell Sage Foundation.

Hemphill, F., Vanneman, A., & Rahman, T. (2011). *How Hispanic and White students in public schools perform in mathematics and reading on the national assessment of educational progress* (No. 2011–459). NCES Report.

Hill, N. E. (2001). Parenting and academic socialization as they relate to school readiness: The roles of ethnicity and family income. *Journal of Educational Psychology, 93,* 686–697. http://dx.doi.org/10.1037/0022-0663.93.4.686

Hill, N. E., Bush, K. R., & Roosa, M. W. (2003). Parenting and family socialization strategies and children's mental health: Low-income Mexican-American and Euro-American mothers and children. *Child Development, 74,* 189–204. http://dx.doi.org/10.1111/1467-8624.t01-1-00530

Hillstrom, K. A. (2009). Are acculturation and parenting styles related to academic achievement among Latino students? (Unpublished dissertation). University of Southern California, Los Angeles, California.

Hong, S., & Ho, H. Z. (2005). Direct and indirect longitudinal effects of parental involvement on student achievement: Second-order latent growth modeling across ethnic groups. *Journal of Educational Psychology, 97,* 32–42. http://dx.doi.org/10.1037/0022-0663.97.1.32

Hong, S., Malik, M. L., & Lee, M. K. (2003). Testing configural, metric, scalar, and latent mean invariance across genders in sociotropy and autonomy using a non-Western sample. *Educational and Psychological Measurement, 63,* 636–654. http://dx.doi.org/10.1177/0013164403251332

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly, 23,* 27–50. http://dx.doi.org/10.1016/j.ecresq.2007.05.002

Hughes, D., Rodriguez, J., Smith, E. P., Johnson, D. J., Stevenson, H. C., & Spicer, P. (2006). Parents' ethnic-racial socialization practices: A review of research and directions for future study. *Developmental Psychology, 42,* 747–770. http://dx.doi.org/10.1037/0012-1649.42.5.747

Hughes, J., & Kwok, O. M. (2007). Influence of student-teacher and parent-teacher relationships on lower achieving readers' engagement and achievement in the primary grades. *Journal of Educational Psychology, 99,* 39–51. http://dx.doi.org/10.1037/0022-0663.99.1.39

Jimerson, S., Egeland, B., Sroufe, L. A., & Carlson, B. (2000). A prospective longitudinal study of high school dropouts examining multiple predictors across development. *Journal of School Psychology, 38,* 525–549. http://dx.doi.org/10.1016/S0022-4405(00)00051-0

Kaufman, A., & Kaufman, N. (2005). *Kaufman Test of Educational Achievement–Brief form manual* (2nd ed.). Circle Pines, MN: AGS Publishing.

Keith, T. Z. (2014). *Multiple regression and beyond: An Introduction to multiple regression and structural equation modeling.* New York, NY: Routledge.

Kline, R. B. (2005). *Principles and practice of structuralequation modeling* (2nd ed.). New York, NY: Guilford Press.

Knight, G. P., Virdin, L. M., & Roosa, M. (1994). Socialization and family correlates of mental health outcomes among Hispanic and Anglo American children: Consideration of cross-ethnic scalar equivalence. *Child Development, 65,* 212–224. http://dx.doi.org/10.2307/1131376

Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies* (Unpublished Report). Los Angeles, CA: California State University. Available from http://www.eric.ed.gov

Lapointe, V. R., Ford, L., & Zumbo, B. D. (2007). Examining the relationship between neighborhood environment and school readiness for kindergarten children. *Early Education and Development, 18,* 473–495. http://dx.doi.org/10.1080/10409280701610846

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist, 35,* 125–141. http://dx.doi.org/10.1207/S15326985EP3502_6

Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school.* Washington, DC: Economic Policy Institute.

Leventhal, T., & Brooks-Gunn, J. (2000). The neighborhoods they live in: The effects of neighborhood residence on child and adolescent outcomes. *Psychological Bulletin, 126,* 309–337. http://dx.doi.org/10.1037/0033-2909.126.2.309

Llagas, C. (2003). *Status and trends in the education of Hispanics (NCES 2003–008). U.S. Department of Education, National Center for Education Statistics.* Washington, DC: U.S. Government Printing Office.

Ma, X., & Klinger, D. A. (2000). Hierarchical linear modeling of student and school effects on academic achievement. *Canadian Journal of Education, 25,* 41–55. http://dx.doi.org/10.2307/1585867

Mardell-Czudnowski, C., & Goldenberg, D. S. (1998). *Developmental indicators for the assessment of learning.* Monterey, CA: American Guidance Service.

McEvoy, A., & Welker, R. (2000). Antisocial behavior, academic failure, and school climate: A critical review. *Journal of Emotional and Behavioral Disorders, 8,* 130–140. http://dx.doi.org/10.1177/106342660000800301

Moon, S. S., Kang, S. Y., & An, S. (2009). Predictors of Immigrant children's school achievement: A comparative study. *Journal of Research in Childhood Education, 23,* 278–289. http://dx.doi.org/10.1080/02568540909594661

National Institute of Child Health and Human Development Early Child Care Research Network. (2000). The relation of child care to cognitive and language development. *Child Development, 71,* 960–980. http://dx.doi.org/10.1111/1467-8624.00202

Oades-Sese, G. V., Esquivel, G. B., Kaliski, P. K., & Maniatis, L. (2011). A longitudinal study of the social and academic competence of economically disadvantaged bilingual preschool children. *Developmental Psychology, 47,* 747–764. http://dx.doi.org/10.1037/a0021380

Olivari, M. G., Tagliabue, S., & Confalonieri, E. (2013). Parenting style and dimensions questionnaire: A review of reliability and validity. *Marriage & Family Review, 49,* 465–490. http://dx.doi.org/10.1080/01494929.2013.770812

Pong, S. L., Hao, L., & Gardner, E. (2005). The roles of parenting styles and social capital in the school performance of immigrant Asian and Hispanic adolescents. *Social Science Quarterly, 86,* 928–950. http://dx.doi.org/10.1111/j.0038-4941.2005.00364.x

Reardon, S., & Galindo, C. (2009). The Hispanic-White achievement gap in math and reading in the elementary grades. *American Educational Research Journal, 46,* 853–891. http://dx.doi.org/10.3102/0002831209333184

Reynolds, C. R., & Kamphaus, R. W. (2004). *BASC-2: Behavior Assessment System for Children* (2nd ed.). Circle Pines, MN: AGS Publishing.

Rhoades, B. L., Warren, H. K., Domitrovich, C. E., & Greenberg, M. T. (2011). Examining the link between preschool social–emotional competence and first grade academic achievement: The role of attention skills. *Early Childhood Research Quarterly, 26,* 182–191. http://dx.doi.org/10.1016/j.ecresq.2010.07.003

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73,* 417–458. http://dx.doi.org/10.1111/j.1468-0262.2005.00584.x

Robinson, C. C., Mandleco, B., Olsen, S. F., & Hart, C. H. (1995). Authoritative, authoritarian, and permissive parenting practices: Development of a new measure. *Psychological Reports, 77,* 819–830. http://dx.doi.org/10.2466/pr0.1995.77.3.819

Roderick, M. (1994). Grade retention and school dropout: Investigating the association. *American Educational Research Journal, 31,* 729–759. http://dx.doi.org/10.3102/00028312031004729

Rodriguez, B. L., & Olswang, L. B. (2003). Mexican-American and Anglo-American mothers' beliefs and values about child rearing, education, and language impairment. *American Journal of Speech-Language Pathology, 12,* 452–462. http://dx.doi.org/10.1044/1058-0360(2003/091)

Sáenz, V. B., & Ponjuan, L. (2011). *Men of color: Ensuring the academic success of Latino males in higher education.* Washington, DC: Institute for Higher Education Policy.

Steinberg, L., Lamborn, S. D., Dornbusch, S. M., & Darling, N. (1992). Impact of parenting practices on adolescent achievement: Authoritative parenting, school involvement, and encouragement to succeed. *Child Development, 63,* 1266–1281. http://dx.doi.org/10.2307/1131532

Suárez-Orozco, M. M., Suárez-Orozco, C., & Qin-Hilliard, D. (2005). *The new immigration: An interdisciplinary reader.* New York, NY: Routledge.

Villavicencio, A., Bhattacharya, D., & Guidry, B. (2013). *Moving the needle: Exploring key levers to boost college readiness among Black and Latino males in New York City.* The Research Alliance for New York City Schools.

Walk, R. A. (2005). *Evaluating the predictive validity of the Speed DIAL version of the DIAL-3, Developmental Indicators for the Assessment of Learning* (Doctoral dissertation). East Tennessee State University, United States. Retrieved from Dissertations & Theses: Full Text database.

Wentzel, K. R. (1993). Does being good make the grade? Social behavior and academic competence in middle school. *Journal of Educational Psychology, 85,* 357–364. http://dx.doi.org/10.1037/0022-0663.85.2.357

Williams, L. S., Alvarez, S. D., & Hauck, K. S. A. (2002). My name is not Maria: Young Latinas seeking home in the heartland. *Social Problems, 49,* 563–584. http://dx.doi.org/10.1525/sp.2002.49.4.563

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management, 27,* 122–154. http://dx.doi.org/10.1002/pam.20310

Yowell, C. M. (2000). Possible selves and future orientation: Exploring hopes and fears of Latino boys and girls. *The Journal of Early Adolescence, 20,* 245–280. http://dx.doi.org/10.1177/0272431600020003001

Zea, M. C., Asner-Self, K. K., Birman, D., & Buki, L. P. (2003). The abbreviated multidimensional acculturation scale: Empirical validation with two Latino/Latina samples. *Cultural Diversity and Ethnic Minority Psychology, 9,* 107–126. http://dx.doi.org/10.1037/1099-9809.9.2.107

Appendix

Standardized Direct, Indirect, and Total Effects on Academic Achievement at the End of First Grade by Ethnic Group

	MA children			DA children		
Variable	Direct effect	Indirect effect	Total effect	Direct effect	Indirect effect	Total effect
1. U.S. identity	—	−.01	−.01	—	.00	.00
2. Ethnic identity	—	.00	.00	—	−.00	−.00
3. U.S. cultural knowledge	—	.01	.01	—	.01	.01
4. Ethnic cultural knowledge	—	.02	.02	—	.01	.01
5. Socialization of independence	—	.12	.12	—	.07	.07
6. Socialization of *respeto*	—	−.06	−.06	—	−.02	−.02
7. Authoritative practices	—	.16	.16	—	.09	.09
8. Authoritarian practices	—	−.19	−.19	—	−.06	−.06
9. Problem behaviors	−.11	—	−.11	−.09	—	−.09
10. Adaptive behaviors	.18	—	.18	.23	—	.23
11. Academic readiness	.93	—	.93	.86	—	.86

Received May 24, 2016
Revision received January 9, 2017
Accepted January 9, 2017 ■

Journal of Educational Psychology
2018, Vol. 110, No. 1, 133–146

0022-0663/18/$12.00 http://dx.doi.org/10.1037/edu0000202

School Readiness Amongst Urban Canadian Families: Risk Profiles and Family Mediation

Dillon T. Browne, Mark Wade, Heather Prime, and Jennifer M. Jenkins
University of Toronto

There is an ongoing need for literature that identifies the effects of broad contextual risk on school readiness outcomes via family mediating mechanisms. This is especially true amongst diverse and urban samples characterized by variability in immigration history. To address this limitation, family profiles of sociodemographic and contextual risk were identified when children were 2 months of age. Subsequently, their indirect effect on school readiness at 4.5 years was evaluated via family investments and maternal responsivity ($N = 501$ families). A latent class analysis yielded four distinct family risk profiles: low socioeconomic status (SES) multilevel risk, 12.0% of sample; maternal abuse history, 15.6%; low-SES immigrant risk, 27.7%; and low risk, 44.7%. Path analyses revealed that children in the low-SES multilevel risk and low-SES immigrant risk profiles had the poorest outcomes in all domains and these effects operated equally and indirectly via investments and responsivity. To date, several studies have suggested that sociodemographic risks impact cognitive outcomes primarily via the investment pathway. The present findings suggest that family relations are equally important when operationalized as observed responsive parenting. Furthermore, pathways of influence are similarly operative despite different patterning of adversity for high-risk immigrant and native born families.

Educational Impact and Implications Statement

Children who grow up in settings of social and economic disadvantage are often behind their peers in learning outcomes by the time they begin school. One reason for this is that families struggle to provide their children with enrichment and learning opportunities during the early years. Another reason is that parents become stressed and struggle to create a household environment—in terms of relationships—that promotes learning. The present study found that this pattern holds for many urban immigrants in Canada, whereby families are struggling economically, parents are less able to promote early learning, and their children are behind by the time they enter school.

Keywords: school readiness, risk, parenting, investments, immigrants

Supplemental materials: http://dx.doi.org/10.1037/edu0000202.supp

The pathways between contextual risk and school readiness are suggested to operate via family experience (Duncan & Magnuson, 2012; Janus & Duku, 2007; Koury & Votruba-Drzal, 2014). These mechanisms include (a) *family investments*, such as resource-based and material investments in child development, in addition to the physical quality of the home; and (b) *family process*, including stress-influenced parenting practices and interpersonal socialization experiences (Bradley & Corwyn, 2002; Conger, Conger, & Martin, 2010). Several studies suggest that investments are a stronger mediator than family process for cognitive domains of school readiness (Guo & Harris, 2000; Linver, Brooks-Gunn, & Kohen, 2002; Yeung, Linver, & Brooks-Gunn, 2002), though this is based on operationalizing family processes as general parenting practices or styles. Studies connecting observed *maternal responsivity* with cognition, as well as social disadvantage, suggest that family processes also serve as a mediator (Mistry, Biesanz, Chien, Howes, & Benner, 2008; National Institute of Child Health and Human Development Early Child Care Research Network, 2005; Raviv, Kessenich, & Morrison, 2004).

Currently, there is a paucity of literature examining the family mediation of broad contextual risk on multiple domains of school readiness (cf. Vernon-Feagans, 2013), especially among urban, multicultural, and immigrant samples. Indeed, there are increasing calls for the integration of risk surrounding migration and resettlement into models of public health and human development (Zimmerman, Kiss, & Hossain, 2011), including the factors that predict the success of children approaching school entry (Boivin & Bierman, 2014). Immigrant families face unique contextual stressors and challenges following relocation, especially socioeconomic viability (Gazso & Waldron, 2009; Koury & Votruba-Drzal, 2014; Perreira, Chapman, & Stein, 2006). Accordingly, person-centered data analytic techniques have been utilized to identify risk patterning across diverse samples (e.g., Pratt, McClelland, Swanson, & Lipscomb, 2016), though samples are generally drawn from

This article was published Online First April 10, 2017.

Dillon T. Browne, Mark Wade, Heather Prime, and Jennifer M. Jenkins, Department of Applied Psychology and Human Development, University of Toronto.

Correspondence concerning this article should be addressed to Jennifer M. Jenkins, Department of Applied Psychology and Human Development, University of Toronto, 252 Bloor Street West, Toronto, ON M5S 1V6, Canada. E-mail: jenny.jenkins@utoronto.ca

the United States and diversity mostly constitutes Latinos and African Americans. The immigration landscape in the 21st century is increasingly global, with more families from Asia and the Middle East immigrating to Western Europe and Canada (Alba & Foner, 2015), and an unprecedented number of children and families are currently living displaced (UNHCR, 2015). Patterns of risk and school readiness remains unexplored in such families, in addition to the mechanisms through which risk exerts influence on development. Accordingly, the present study employs a unique combination of person-centered and variable-centered approaches to answer two unanswered questions: (a) *Who* are the families that are most at risk for poor school readiness among a diverse sample of Canadian immigrants and nonimmigrants? and (b) *How* does risk impact school readiness via the family?

Social Disadvantage, Cognitive Outcomes, and School Readiness

Cognitive functioning at school entry is as a keystone predictor of developmental health across the life course (Boivin & Bierman, 2014). Accordingly, the importance of identifying pathways toward *school readiness* has been emphasized, referring to a multidimensional set of cognitive and socioemotional skills predicting school success. These include the interconnected abilities of language, executive functioning (EF), theory of mind (ToM), and early academics (Boivin & Bierman, 2014; Duncan et al., 2007; Janus & Duku, 2007). To date, deficits in EF and language have been most strongly tied to socioeconomic status (SES) variation (Hart & Risley, 1992; Noble, Norman, & Farah, 2005). However, difficulties in ToM are also related to individual variation in EF and language (Astington & Jenkins, 1999; Hughes, 1998). It is important to understand etiological sources of variation in school readiness, given that disparities amplify over time, contributing to social discrepancies in educational success (Sirin, 2005). For example, using six longitudinal data sets, Duncan and colleagues (2007) demonstrated that academic ability and attention at school entry provided the strongest prediction of later achievement. Such findings highlight the importance of exploring the mechanisms through which cognitive variability emerges before children enter school.

Mechanisms of Social Disadvantage

Evidence citing the importance of monetarily based material investments in cognitive development is sizable and available elsewhere (Conger et al., 2010; Guo & Harris, 2000; Mistry et al., 2008; Yeung et al., 2002). Families with greater income (or disadvantaged families who receive income supplements) can provide their children with more opportunities that promote the development of their human skills, including school readiness (Heckman, 2006). This mechanism has been called resources, human capital, or investments. We refer to this pathway using the *family investments* label. The other mediating mechanism concerns the impact of economic stress on nonfinancial aspects of family life, such as the adjustment of caregivers, and the consequences of this for parent–child interaction quality. Previous literature has described this pathway as socialization, family stress, the psychological model, or family process. We adopt the *family process* label.

The role of family process (i.e., parenting) as a relational predictor of school readiness remains surprisingly equivocal. This is likely because of differences in parenting measurement. For example, using brief interviewer-rated parenting items from the HOME Inventory (Bradley, 1994), Yeung and colleagues (2002) found that parenting did not mediate the impact of SES on academic achievement over and above family investments. Guo and Harris (2000) found significant but small effects for parenting on child cognition, but concluded that family investments were the most influential. Studies that assess other aspects of parenting (authoritative and authoritarian rearing) have similarly failed to find a mediating role of parenting (Linver et al., 2002). Conversely, studies that examine *maternal responsivity* during observed and coded interactions find a strong relationship between parenting and cognitive outcomes, with evidence of mediation (Mistry, Benner, Biesanz, Clark, & Howes, 2010; Mistry et al., 2008; National Institute of Child Health and Human Development Early Child Care Research Network, 2005; Vernon-Feagans, 2013).

Several subcomponents of maternal responsivity have been emphasized, and there appears to be phenomenological convergence among these constituent parts (de Rosnay & Murray, 2012; Meunier, Boyle, O'Connor, & Jenkins, 2013). First, *sensitivity* refers to the decoding of children's cues for the purpose of responding in a contingent fashion, in addition to the presence of affective warmth during interactions (Claussen & Crittenden, 2000). In the present investigation, aspects of sensitivity included responsiveness to the child's verbal and nonverbal overtures, child-focused engagement, child mindedness, and responsive facilitation. Second, *mutuality* describes the synchrony and quality of dyadic interactions, including positive shared attention and affect, turn taking, and conversational fluidity (Ensor, Spencer, & Hughes, 2011). Finally, *positive control* describes parental bids for control and the manner in which parents shape child behavior through specific praise, explanation, elaboration, suggestion, and open-ended questioning (Deater-Deckard et al., 2001; Deater-Deckard, Pylas, & Petrill, 1997). There is overlap between this aspect of responsivity and "autonomy-support" or "scaffolding" (Bernier, Carlson, Deschênes, & Matte-Gagné, 2012; Bernier, Carlson, & Whipple, 2010).

A comprehensive understanding of the family process pathway requires theoretical linkages between responsive caregiving and cognitive development. Most of these accounts are consistent with models suggesting that psychological ontogenesis is a product of internalized interpersonal experience (Fernyhough, 2008; Vygotsky, 1978). That is, responsive maternal behavior during interactions facilitates the coactivation of intersubjective processes that serve as precursors to autonomous cognitive abilities (Blair & Raver, 2012). Across repeated experience, these dyadic phenomena become increasingly represented within individual cognition. The internalization of interpersonal experience has been called *experiential canalization*, defined as the experience-dependent "selective optimization" of particular cognitive functions based on certain rearing environments (Blair & Raver, 2012). For example, Bernier and colleagues (2010, 2012) have shown that the emergence of EF reflects a gradual internalization of coregulated parent–child interactions. Moreover, there is interdependence among EF capacity and the development of math and reading, language, and ToM (Benson, Sabbagh, Carlson, & Zelazo, 2013; Fitzpatrick, McKinnon, Blair, & Willoughby, 2014). That is, high-quality dyadic exchanges seem to promote the emergence of a cognitive set that facilitates self-directed learning and successful interpersonal exchanges by the time of school entry.

Considering Social Disadvantage From a Person-Centered Perspective

Studies of social disadvantage often employ univariate economic indices as a proxy for environmental adversity, including family income, education, or income-to-needs ratios. Another approach is the *cumulative risk* perspective, which suggests that the aggregation of risks contributes to developmental consequences as opposed to their presence in isolation (Evans & Kim, 2012; Vernon-Feagans, 2013). Although there is substantial evidence to support these methodologies, there are limitations to variable-centered approaches. Relying on basic metrics or simply counting risks can obscure the existence of subtypes of social disadvantage (Pratt et al., 2016). This may be especially true among heterogeneous groups, characterized by diversity in terms of immigration history, refugee status, life history, ethnicity, and culture (Gazso & Waldron, 2009; Koury & Votruba-Drzal, 2014; Perreira et al., 2006). Using person-centered approaches, a few researchers have suggested that such risks may not simply accumulate, but may combine in prototypical ways to create discrete *family risk profiles* that are differentially related to developmental outcomes (e.g., Brody et al., 2013; Copeland, Shanahan, Costello, & Angold, 2009; Lanza, Rhoades, Nix, Greenberg, & Conduct Problems Prevention Research Group, 2010). Such methodology has public health relevance, as analyses can help identify and target homogenous subgroups of persons who are differentially at risk based on readily identifiable characteristics. Subsequently, initiatives for prevention and intervention can be tailored to the particular strengths and challenges for the subgroups in question (Lanza & Rhoades, 2013).

At present, it is unclear (a) how risk factors for poor school readiness combine for families characterized by diversity in immigration history, and (b) how these profiles exert influence on children via family investments and family process. This is a limitation despite an increasingly mobile global population, characterized by growing diversity and economic challenge; families often settle in Europe and Canada in addition to the United States, and there are marked differences in immigration policy across these developed nations (Alba & Foner, 2015), though much research does not reflect this reality. For example, two thirds of Canadian immigrants are from economic categories (such as entrepreneurs and skilled laborers) compared with only 16% in the United States (CIC News, 2015). Exploring patterns of human development and health among immigrant families in multiple societies is a global public health concern (Zimmerman et al., 2011). Resettled households are faced with a number of struggles. Central among these may be socioeconomic viability, given that immigrants are often of higher education (e.g., because of point systems) but do not have their credentials validated in host societies, resulting in economic hardship and underemployment (Picot, 2004). Furthermore, economic challenges appear to exacerbate when individuals are of minority status (Gazso & Waldron, 2009; Creese & Wiebe, 2012) and parents (Jansen et al., 2010; Perreira et al., 2006). Additional difficulties include the negotiation and acquisition of work visas, employment, housing, and fiscal documentation, and limited social capital/support. Consequently, many immigrant families congregate in low-income neighborhoods along with the native poor, where there is higher crime, deviance, noise, pollution, and, ultimately, stress. Unfortunately, the consequences of these phenomena on family life and school readiness remain unclear.

Current Study

The present study explored two central research questions: (a) *Who* are the families that are most at risk for poor school readiness among a sample of ethnically diverse Canadian immigrants and nonimmigrants? and (b) *How* does risk operate via family investment and family process among these identified families? Given the aforementioned challenges associated with immigrant status, we hypothesized a unique group of immigrants to emerge who are most notably struggling in the economic arena. We expected observable consequences on school readiness via the family investment and process pathway for these families, given that economic stressors are central to the causal mechanisms outlined by these models. Consistent with the cumulative risk perspective, we also expected to find a group of families who had risk in multiple domains, with corresponding consequences in family investments and family process (i.e., responsivity) pathways.

Method

Participants

Participants are derived from the intensive sample of the Kids, Families, and Places (KFP) Study. Women giving birth in Toronto and Hamilton, Ontario, between April 2006 and September 2007 were considered for participation. Families were recruited through Healthy Babies Healthy Children, a program in which parents of all newborns are contacted within a week of their birth. Inclusion criteria included the presence of an English-speaking mother, agreeing to be filmed in the home, the presence of at least two children who are less than 4 years, and a newborn greater than 1,500 g (>3.30 lbs, i.e., not very low birth weight [VLBW]). VLBW children were excluded, given that this is a medically complex group, presenting with a number of neurodevelopmental concerns, representing a small proportion of births (1.3%; Kowlessar, Jiang, & Steiner, 2013). As the KFP study is an epidemiological and prospective birth cohort concerned with multiple psychosocial and genetic influences on human development, this group was omitted so as not to bias general patterns of association in the sample. Multiparous mothers were exclusively recruited, given that the KFP Study was also concerned with exploring development through within-family differences. Only the target (youngest) child was included in the current study in order to track children from birth to school entry. Reasons for nonenlistment included inability to contact and refusals. The University of Toronto Research Ethics Board approved all procedures, including informed consent. We compared our sample (N = 501) with the general population of Toronto and Hamilton using 2006 census data, limiting the census to women between 20 and 50 years and having at least one child. Families were compared on immigrant status, number of persons in the home, family type, maternal income, and education. Families were of similar size (M = 4.52, SD = 1.01 vs. M = 4.13, SD = 1.22) and income (median

C$30,000–$39,999 vs. census population M = C$30,504.16, SD = C$37,808.12). As our sample was recruited shortly after childbirth, there are fewer nonintact families than in the population (5% vs. 16.8% lone-parent families; 4.3% vs. 10.3% stepfamilies). The ratio of Canadian born to immigrants was somewhat higher in the current sample (57.7% vs. 47.6%), likely because of an English language requirement for participation. Also, more study mothers had earned a bachelor's degree or higher (53.3% vs. 30.6%). Of participating mothers, 56.5% self-identified as being of European descent, 14.6% as South Asian, 9.3% as Black, 12% as East Asian, and 8.6% as other. Consistent with the Canadian cultural mosaic, immigrant caregivers (n = 233) reported being from any of 19 countries plus a residual category. The most common countries were China (10.3%), Sri Lanka (9.9%), India (9.0%), Philippines (7.7%), Pakistan (5.6%), United States (3.4%), and Guyana (2.6%). Of immigrant caregivers, the average amount of time in Canada was 10.13 years (SD = 8.96).

Of participating families, 74.1% of families were two-child families, 18.8% were three-child families, and the remaining 7.2% had four or more children. Mean child age at baseline was 2.0 months (SD = 1.06), and 49.3% of children were female. Families were initially followed up when children were 1.6 years (SD = .16), when 397 (79.2%) families remained, and again when children were at the age of school entry (4.79 years old, SD = .28), when 323 (64.5%) families remained. For simplicity, these time points are referred to as birth, 18 months, and school entry/4.5 years. Retained families were somewhat higher functioning, as is often the case in longitudinal research (see Missing Data section). Best practices were followed for missing data (i.e., multiple imputation across 25 data sets) in order to minimize the effect of attrition (Graham, 2009).

Measurement

Distal risk factors. Risk measures are described in brief and comprehensive descriptions are available in the Appendix in the online supplemental materials. This is because of the large number of distal risks examined in the present study. Risks factors were assessed at birth in the domains of early maternal history (e.g., witnessing or experiencing abuse, familial history of psychopathology and substance abuse), current caregiver factors and household risk (e.g., immigration status, maternal depression, marital conflict, single-parenthood), socioeconomic risk (e.g., income under $20,000, absence of postsecondary education, recipient of income supplement), and neighborhood risk (e.g., poor observed neighborhood quality, victimization, high concentration of low-income housing). Measures were predominantly maternal report, though home visitors made reliable neighborhood observations based on the Block Environment Inventory (Cronbach's alpha [α] = .96; McGuire, 1997) as detailed in the Appendix. Additionally, families were linked to the Statistics Canada (2006) Canadian National Census using the first three characters of their postal code, providing additional information (e.g., census tract poverty and unemployment rates). Thirty-two risks across these domains were coded as being either present (1) or absent (0) based upon naturally occurring criteria (e.g., teenage parenthood) or distributional cutoffs (i.e., +1 standard deviation toward the risky or adverse pole). More information is provided in the online supplemental materials.

Family mediating variables. Mediating variables were assessed during home visits at 18 months. The family investment pathway was assessed as quality of the home environment and material investments, indexed using the observer-response items of the HOME Inventory. The family process pathway was measured as maternal responsivity during videotaped mother–child interactions (henceforth referred to as the "maternal responsivity pathway"). Trained interviewers used an adapted version of the HOME scale (Bradley, 1994). The standard HOME scale is a mixture of observer and parental report. We used only the observer items in order to avoid parental response bias and to avoid contamination among assessments of the process and investment pathways. In total, observers rated 10 items on a 3-point scale assessing cleanliness/safety of physical environment and toys/educational-materials available to the children. Negatively worded items were reverse coded and individual items were averaged to create a composite reflecting family investment (Cronbach's α = .68).

Also at this time, mothers were videotaped interacting with the target child for 15 min. There were three 5-min tasks. First, there was a *free play with no toys*, in which mothers were instructed to play with children as they normally would but without any materials. Second, there was a *structured teaching with toys*, in which dyads were given a pegboard with circles and squares of different colors and instructed to copy a picture. Specifically, mothers were asked to teach their children how to construct the pattern in the picture, in which the pattern was intentionally beyond the child's developmental level in order to elicit maternal teaching. Finally, there was a *reading task*, during which the mother was asked to make up a story to a wordless picture book. These tasks were selected in order to assess mother's capacity to engage the child positively during common tasks of early childhood and to challenge the child's attention and self-regulatory abilities.

Maternal responsivity was assessed using the Sensitive Responding and Mutuality scales of the Coding of Attachment Related Parenting scheme (Matias, Scott, & O'Connor, 2006) and the Positive Control scale of the Parent–Child Interaction System (Deater-Deckard et al., 1997). Sensitive responding measures the ability of mothers to display awareness of their child's needs, to be sensitive to the child's signals, and demonstrate perspective taking from the child's vantage point. Mutuality is a dyadic code that reflects conversational reciprocity, sharing of affect, joint engagement during tasks, and open physical posture. The Positive Control scale assesses positive aspects of a mother's style of directing or influencing child behavior, including praise and open-ended questions. A composite was computed by averaging the Sensitive Responding, Mutuality, and Positive Control subscales across all tasks. A senior graduate student, the "expert rater," trained a group of research assistants on the coding scheme. Once they were trained to criterion (i.e., Cronbach's $\alpha > .80$; Stemler, 2004), they independently coded videos. Following data submission for the calculation of reliability which occurred throughout the coding period (every 10th tape or weekly), coder discrepancies ($< .80$) were discussed. Internal consistency (Cronbach's α = .85) and interrater reliability (Cronbach's α = .94) were good.

Child cognitive outcome variables. Cognitive outcomes were assessed at the age of school entry. *Receptive vocabulary* was assessed with the Peabody Picture Vocabulary Test—Fourth Edition (PPVT-IV; L. M. Dunn & Dunn, 2007). An examiner presents a series of stimulus cards with four images and reads a word aloud.

The child is required to indicate the picture that corresponds best to the target word. The PPVT-IV is a norm-referenced test and was validated on a representative sample. ToM was measured using an adaptation of the scale described by Wellman and Liu (2004). The first three tasks assessed children's understanding of diverse desires and beliefs, and knowledge and ignorance, followed by tasks that assessed more sophisticated ToM understanding, such as false belief, belief-based emotion, and real-apparent emotion. If children failed two consecutive tasks, testing was stopped. For all ToM tasks, stories were enacted with the use of puppets and props. A total score across all tasks (pass or fail) was computed. Internal consistency was good (Cronbach's α = .87).

EF was assessed using the Bear–Dragon task (Reed, Pien, & Rothbart, 1984) and the Dimensional Change Card Sort (DCCS; Zelazo, 2006). For the Bear–Dragon task, children were instructed to do what they were told by the nice bear (e.g., "touch your nose"), but not to do what they were told by the mean dragon. Children were scored for total number of correct responses (0–10) on five dragon and five bear trails. For the DCCS, children were required to sort a series of bivalent test cards, first according to one dimension (e.g., color), and then according to the other (e.g., shape). Children who pass the post-switch phase of the standard version of the DCCS may proceed immediately to the border version, which uses the same target cards as the standard version. The border version consists of 12 trials. Children are required to sort cards based on "border" criteria ("If there's a border, play the color game. If there's no border, play the shape game"). Both tasks tap into set shifting, working memory, and inhibitory control, though certain aspects are emphasized depending on the task (i.e., the Bear–Dragon task emphasizes inhibitory control, and DCCS emphasizes working memory and set shifting). Consistent with previous studies (Bernier et al., 2010; Carlson, Moses, & Claxton, 2004; Hughes & Ensor, 2005), both measures were correlated (r = .34, $p < .001$; medium effect size) and combined into a composite by z scoring and averaging across tasks. (Note that analyses were also conducted using *either* the Bear–Dragon or DCCS, and patterns of association were identical).

Finally, academic skills were assessed using three subtests from the Woodcock-Johnson Third Edition—Tests of Educational Achievement (WJIIIACH; Woodcock, McGrew, & Mather, 2001). Age equivalents were combined into a composite score. *Print and sight-word recognition* was assessed using the Letter-Word Identification subtest, requiring children to name letters and words aloud. *Reading comprehension* was assessed with the Passage Comprehension subtest, requiring children to read a partially incomplete passage and identify the missing word. Finally, *numeracy and mathematical ability* was assessed using the Applied Problems subtest. Children responded to math problems that were read aloud. Some items had visual stimuli. The WJIIIACH is widely used in research and clinical settings, as it provides reliable, valid, and brief assessment of early academic functioning. Similar to the PPVT-IV, the WJIIIACH was normed on a large, representative, and diverse sample. There was significant and good convergence among the three academic subtests (intraclass correlation coefficient = .63).

Covariates. Covariates were selected based upon their relation with cognitive outcomes, including child birth weight (Landry, Smith, & Swank, 2006), maternal language (Hoff, 2003), and gender (Zambrana, Ystrom, & Pons, 2012). Thus, analyses adjusted for child gender, whether or not English was the primary language spoken in the home, birth weight, and age at follow-up in years.

Analysis

Analyses were conducted in two steps: (a) latent class analysis (LCA) identifying profiles of distal risk at birth, and (b) path analysis identifying the effects of risk profile on child outcomes at school entry via family mediators (family investments and maternal responsivity) at 18 months. First, the 32 risk factors across domains of adversity at birth (maternal history, current caregiver and household characteristics, SES, and neighborhood risk) were recoded into being either present (1) or absent (0) in accordance with criteria described previously and outlined in the Appendix found in the online supplemental materials. These 32 risk variables were subjected to LCA, in which solutions fitting between two and six classes were compared. Guidelines for selecting the best-fitting number of classes were followed (Asparouhov & Muthén, 2012). When model fit is optimal and the entropy statistic is approaching 1, there is little ambiguity in class membership. Thus, participants were assigned to the latent class to which they had the highest probability of membership. The resulting nominal class membership variable was recoded into a series of dummy variables, and the class with the lowest probability of endorsing overall distal risk (i.e., low-risk group) was treated as the reference category. These dummies served as the global indicators of distal risk, which were hypothesized to impact school readiness via family mediators at 18 months. To evaluate this, a path model was constructed that (a) regressed mediators onto the class dummies, and (b) regressed cognitive outcomes onto mediators and class dummies. All pathways controlled for covariates, and a maximum likelihood estimator was used that is robust to nonnormality (Muthén & Muthén, 1998–2010). Indirect effects were independently evaluated for each class using the Sobel method (Sobel, 1987). Partially standardized indirect effect sizes (ab_{ps}) are reported, taken as the product of the unstandardized indirect pathways divided by the standard deviation of the outcome variable ($a{*}b/\sigma_Y$; Preacher & Kelley, 2011). This quantifies the relative difference in child outcomes in standard deviation units for a one-unit increase in the dummy-variable risk indicator (i.e., being in a risk group, relative to the low-risk group), operating indirectly via the family mediators. Finally, model constraints and Wald tests were used to examine the null hypothesis that indirect effect sizes (i.e., pathways via family investment vs. maternal responsivity) for a particular risk group and outcome were equivalent.

Analysis of Attrition and Missing Data Procedures

Of the initial 501 families, 323 (64.5%) remained at follow-up. Retained families were compared with those who dropped out on baseline risk indicators included in the LCA. Mothers lost to follow-up were higher risk in a number of areas: mothers' fathers had a drug or alcohol problem, teenage motherhood, unemployment, incomes less than $20,000, and not owning their home. Both mothers and partners lost to follow-up were significantly less likely to have postsecondary education. The neighborhoods of noncompleters were viewed as significantly less trusting and rated by home visitors to be of lower quality. These neighborhoods also had significantly higher rates of census-assessed lone parenthood and poverty. Notably, there was an equal proportion of immigrants retained (44.6%) versus lost to follow-up (50.0%), $\chi^2(1) = 1.35$, $p = .26$. Although there was no missing data for the LCA conducted at baseline, procedures to handle longitudinal missing data were taken from Graham (2009) in the path analysis. Class membership and a series of auxiliary variables (child gender, age, birth weight, gestational age, ethnicity, language spoken in the home, years of residency in Canada for immigrants) were used in an imputation model that resulted in 25 multiple imputation sets. Path models were run across these 25, and estimates were pooled according to Rubin's Rules.

Results

Family Risk Profiles at 2 Months Identified Through LCA

Model fit is presented in Table 1. A four-class solution was selected as the best fitting model because of improvements in the Akaike, Bayesian, and sample-size adjusted Bayesian information criteria, and a statistically significant Lo-Mendell-Rubin likelihood-ratio test. Entropy in all models was similar. The Bayesian information criterion has its lowest value for the four-class solution; this model may be optimal in terms of balancing parsimony and specificity. Random starts and final stage optimizations were increased, to which the optimal log likelihood was robust. The parametric bootstrapped likelihood ratio test was conducted in order to replicate the solution, confirming the four-class solution over the three-class model, -2^{*}*Loglikelihood* (34) = 219.35, $p < .001$. The final four latent classes are visually depicted in Figure 1. Probability of each risk being present is plotted as a function of latent class membership (i.e., risk profile).

One class had a high probability of endorsing most all risks. This group was referred to as *low-SES multilevel risk* ($n = 60$; 12%). These families had mothers who grew up in risky homes, in which there was parental psychopathology, observed and experienced abuse, and family dissolution. They had the highest probability of being headed by a single-parent mother and highest probability of marital conflict. Thus, when there was a partner present, the relationship tended to be acrimonious, making this a class characterized by hostile and ineffective relationships. Additionally, these families had high probabilities of endorsing economic risk. The second group had elevated probabilities of endorsing risks that were almost exclusively in the maternal history domain. This group was referred to as the *abuse history* group ($n = 78$; 15.6%). Mothers from this group also grew up in risky homes, with parental psychopathology, observed abuse, and family dissolution. Interestingly, mothers in this group were less likely to be a *target* of abuse compared with the low-SES multilevel risk group. This group did not endorse risk in other domains that measure current levels of adversity, such as depression or marital conflict, nor were they of low SES. A third group had mothers with the highest probability of being born outside of Canada. These families were referred to as *low-SES immigrant risk*, as they also had high probabilities of endorsing economic risk ($n = 139$; 27.7%). By examining maternal history of risk, it is clear that these families were not at measureable risk before coming to Canada (i.e., no history of family psychopathology, abuse or family separation). Teenage parenthood, homemaking (maternal unemployment), and larger family size may be culturally normative as well. However, these families were presently living in neighborhoods that were of poorer quality. Finally, there was a group that had low probability of endorsing all risk factors, which was referred to as *low risk* ($n = 224$; 44.7%).

Table 1
Fit Indices for Latent Class Models Examining 32 Risk Indicators as a Function of One Categorical Latent Variable Identifying Risk Profile Membership

# Classes	AIC	BIC	aBIC	Entropy	*p* (LMR)
2	12674.98	12953.27	12743.79	.87	<.01
3	12320.51	12742.17	12424.76	.86	.10
4	**12169.16**	**12734.19**	**12308.86**	**.86**	**.05**
5	12052.79	12761.18	12227.93	.88	.19
6	11954.03	12805.78	12164.62	.89	.21

Note. A four-class solution (in bold) was evaluated to be the best-fitting model. AIC = Akaike information criterion; BIC = Bayesian information criterion; aBIC = sample size adjusted Bayesian information criterion, *p* (LMR) = *p* value of the Lo-Mendell-Rubin adjusted likelihood ratio test for *k* versus *k* – 1 classes.

Impact of Risk Profile on Cognition via Family Investments and Maternal Responsivity

Descriptive statistics are presented in Table 2. Estimation of the path model in Figure 2 permitted tests of mediation hypotheses. Only significant pathways are presented in the figure, though all pathways were estimated and complete output is available in tables (see Table 3, Table 4, the Appendix in the online supplemental materials). That is, the indirect effects of risk profile on cognitive outcomes at school entry were examined via family mediators (family investments and responsivity).

Households characterized by the low-SES immigrant risk and low-SES multilevel risk profiles had significantly lower levels of investments and responsivity relative to the low-risk profile. Additionally, family investments were significantly and positively associated with academic achievement, receptive vocabulary, and ToM, whereas maternal responsivity was significantly and positively associated with all child outcomes. The maternal history of adversity profile did not significantly differ from the low-risk group in terms of mediators or outcomes, with the exception of EF. Interestingly, there was a direct effect whereby children from these households had significantly higher levels of EF. All implied indirect pathways in Figure 2 were statistically significant (see Table 3) and can be interpreted in a similar fashion. For example, membership in the low-SES multilevel risk group versus the low-risk group is associated with ToM scores that are .17 standard deviation units lower via investments, and .15 units lower via responsivity. The Wald test examining the null hypothesis that assumes statistical equivalence across investment and responsivity pathways is not rejected, $\chi^2(1) = .10$, $p = .75$, suggesting that the respective paths are not significantly different in size. Analogous results were obtained for all indirect pathways, in which family investments and responsivity offer significant and similarly small-sized indirect effects. In other words, the effects of membership in the low-SES immigrant and low-SES multilevel risk profiles on child

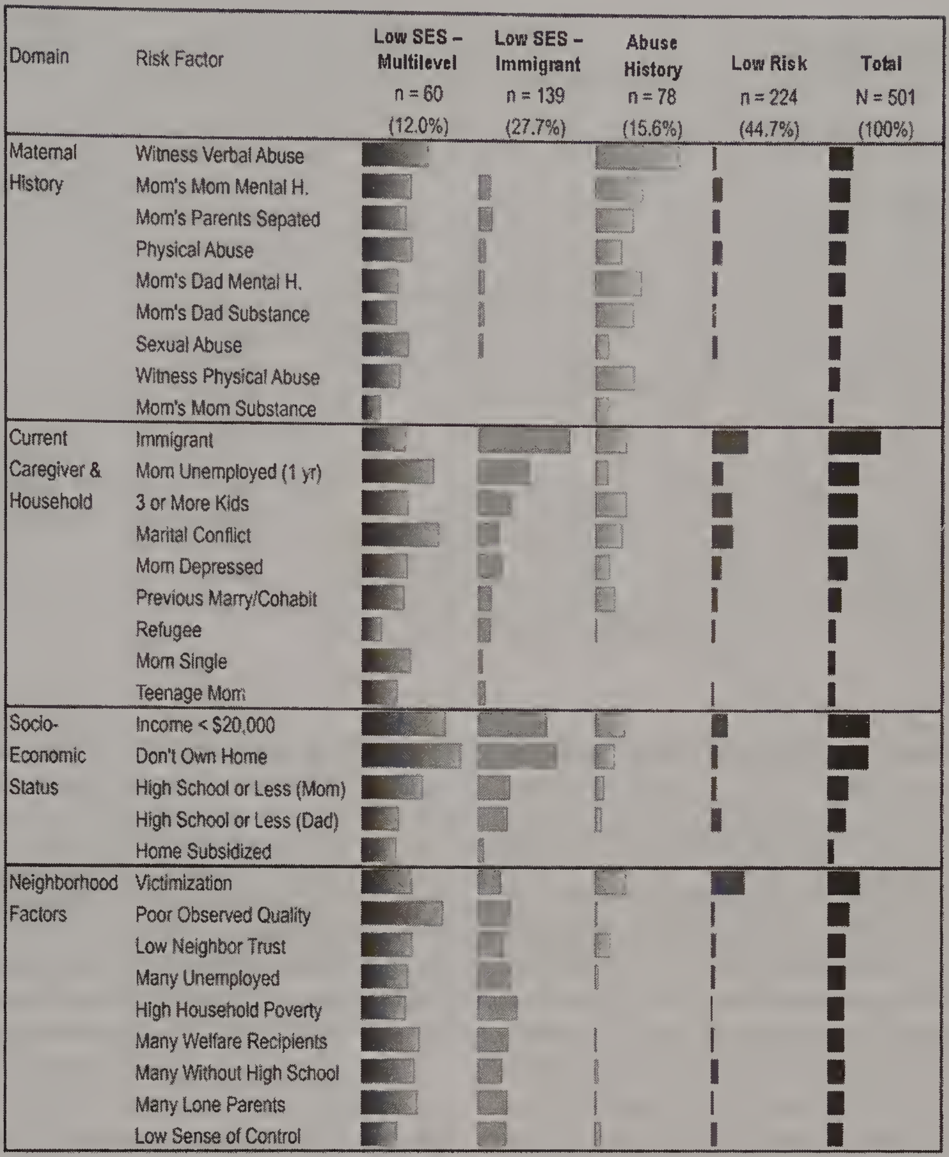


Figure 1. Probability of risk factor by risk profile membership. Each bar represents the probability that an individual in the indicated group endorsed the dichotomized risk factor in question.

outcomes operates indirectly and equally through both family mediators. There is one exception for EF, in which only responsivity was a significant predictor. However, the nonsignificant Wald test indicates that the indirect effects do not significantly differ from one another, likely because of small absolute indirect effect sizes.

As supplementary analyses, the magnitudes of the indirect effects via responsivity or investments were compared across low-SES immigrant risk and low-SES multilevel risk groups. All comparisons were nonsignificant, suggesting the impact of risk via maternal responsivity *and* investments is of similar size for both risk groups. Furthermore, all possible Risk Profile × Maternal Sensitivity and Risk Profile × Family Investment interactions effects were tested in order to determine whether the effects of family mediators on outcomes were similar across risk profiles. None of these effects were statistically significant.

Discussion

To date, there has been a paucity of research examining the mediation of broad contextual risk and school readiness via the family environment, particularly among multicultural, urban, and immigrant samples living outside of the United States. Theoretical models have emphasized two mediating mechanisms: (a) *family investments*, described as material resources for child enrichment and quality of the home environment, and (b) *family processes*, referring to parenting, socialization practices, and supportive and attentive relationships (Bradley & Corwyn, 2002; Conger et al., 2010). The current findings add to a growing body of research suggesting that contextual influences on school readiness operate simultaneously through both mechanisms (Mistry et al., 2008; Raviv et al., 2004; Vernon-Feagans, 2013; Yeung et al., 2002). Moreover, these pathways are similarly operative for low-SES multilevel risk families *and* a previously unidentified group of low-SES Canadian immigrant families. Given the stability of early and later cognitive development (Stanovich, 1986), children in native born and immigrant families who struggle with these skills at school entry are less likely to experience social, educational, economic, and occupational success (Almedom, 2005).

The identification of family subgroups via LCA adds specificity to variable-centered studies that examine family pathways using dimensional metrics of SES or cumulative risk. There is great

Table 2
Descriptive Statistics and Intercorrelations

Variable	1	2	3	4	5	6	7	8	9	10	*M* (*SD*)
1. Girl											.50 (.50)
2. Age (Follow-up)	−.06										4.79 (.23)
3. Birth weight (kg)	−.12**	−.01									3.41 (.50)
4. English	−.02	.10*	.14**								.79 (.41)
5. Investments	.00	−.03	.12**	.16**							2.37 (.33)
6. Responsivity	.15**	.09*	.22**	.16**	.36**						3.50 (.70)
7. Achievement	.01	.27**	.07	.03	.27**	.36**					5.01 (.58)
8. Vocabulary	.03	.09*	.31**	.26**	.39**	.45**	.40**				103.76 (11.80)
9. ToM	.20**	.23**	.09*	.13**	.27**	.39**	.31**	.40**			3.70 (1.26)
10. EF	.03	.25**	.15**	.10*	.22**	.37**	.37**	.49**	.35**		−.01 (.66)
Latent class											
11. Low-SES multilevel	.02	.04	−.07	.16**	−.30**	−.17**	−.11*	−.09*	−.01	−.04	.12 (.33)
12. Low-SES immigrant	.00	−.06	−.10*	−.40**	−.23**	−.18**	−.07	−.26**	−.12**	−.13**	.27 (.44)
13. Abuse history	.02	−.02	.05	.12**	.21**	.13**	.06	.19**	.13**	.16**	.16 (.36)
14. Low risk	−.02	.04	.10*	.16**	.25**	.18**	.09*	.15**	.02	.04	.45 (.50)

Note. ToM = theory of mind; EF = executive functioning; SES = socioeconomic status.
$^{*}p < .05$. $^{**}p < .01$.

utility in the LCA approach in population and public health, as prevention and intervention initiatives can be targeted toward families who present with a unique constellation of needs (Lanza & Rhoades, 2013). For example, the present findings identify difficulties among a unique group of Canadian immigrants. It appears that there is a deleterious impact of SES (and corresponding neighborhood risk) on child cognition via investments and responsivity within urban immigrant households, even when other risks are absent (e.g., maternal depression, divorce, or marital conflict). It is noteworthy that the multilevel risk homes are also of low SES. In other words, it is possible that low SES, per se, is the driving force of distal risk in both of these risk groups. That is, financial stress may be sufficient to activate the ecological transmission of risk via both family investment and family process pathways, even when other risks are absent. Empirical models arguing for the primacy of poverty in the family mediation of contextual risk have been successfully demonstrated (Evans & Kim, 2012). However, if risk in the immigrant homes simply reflected absence of funds, one may expect these households to only display lower levels of family investments (i.e., not maternal responsivity), given that the former pathway directly implicates a family's economic access to goods and services. Stated differently, this would mean that immigrant risk only operated via the family investment pathway and not via the family process pathway. However, consistent with the family stress model (Conger et al., 2010), whereby parenting becomes disrupted by socioeconomic stressors, we see that mothers in risky immigrant families also display lower levels of responsivity. Indeed, there are parenting disruptions despite the fact that there is a low probability of many other psychosocial risks being endorsed.

These findings highlight the importance of considering the economic hardship of immigrant families in Canada, especially when they are minorities (Gazso & Waldron, 2009) and parents (Perreira et al., 2006). Evidence suggests that, despite these challenges, many children of immigrants thrive throughout this remarkable transition and may surpass their host-country counterparts (Coll & Marks, 2012). However, contrary to earlier beliefs, North American immigrants in the 21st century are *not* always catching up economically to the general population (Walks & Bourne, 2006). Such families often end up in neighborhoods that are proximal to the residences of the host-country poor, and fail to experience upward social mobility and resettlement. It is conceivable that functional families that immigrate to North America (in the present study, Canada), who are met with marginalization and socioeconomic immobility, are at the outset of a multigenerational cycle of adversity that undermines school readiness. Furthermore, despite the fact that Canada tends to favor immigrants from economic categories (e.g., entrepreneurs and skilled laborers) compared with the United States (66% vs. 16%), we see a sizable group of immigrants who have children struggling with school readiness, partially because of adversity in the family context. In light of this finding, population-level initiatives targeted at promoting school readiness would be incomplete without careful consideration of the economic success of immigrants, in addition to the impact of SES on the family environment.

Beyond neighborhood stress and the psychological impact of economic challenge, it is possible that other unmeasured psychosocial challenges are influencing parenting for individuals in the low-SES immigrant *and* low-SES multilevel risk groups. For example, the current study did not employ measurement in the domain of parental stress, or stress associated with specific child rearing responsibilities (Abidin, 1990), which is higher in low-SES settings (Pinderhughes, Dodge, Bates, Pettit, & Zelli, 2000). Nevertheless, our findings contribute to family stress theory, suggesting that family process (i.e., responsivity) can be disrupted by economic stress, even in the absence of broadband psychosocial risk. Moreover, such findings echo other studies and theories that emphasize the influence of macroeconomic conditions on family dynamics and developmental health, including recession (Lee, Brooks-Gunn, McLanahan, Notterman, & Garfinkel, 2013) and income inequality (Wilkinson & Pickett, 2009), among immigrants and nonimmigrants, alike.

Given the identification of an immigrant risk group, it is worth mentioning that previous research demonstrates cross-cultural validity in the maternal responsivity construct. Although Western definitions of attachment-related parenting and parenting styles have been called into question, particularly among Eastern and

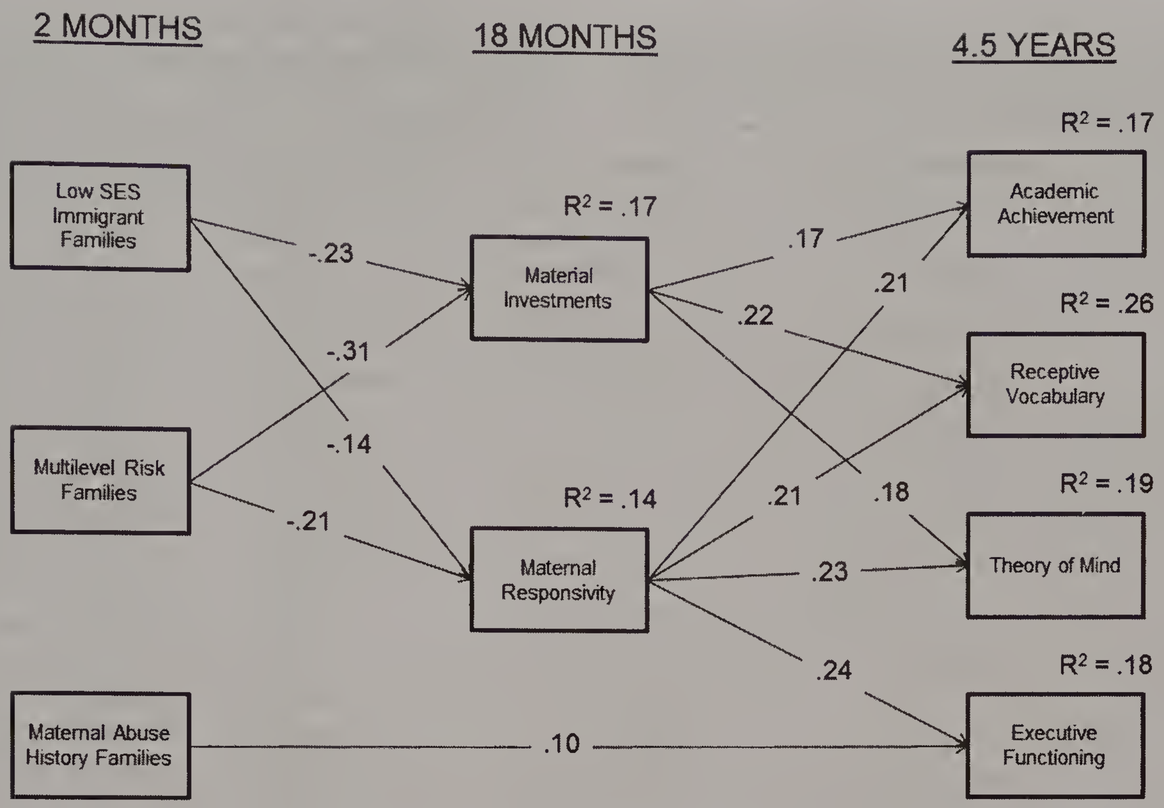


Figure 2. Indirect effects of risk profile membership at 2 months on school readiness outcomes at 4.5 years via family investments and maternal responsivity. Model fit: $\chi^2(3) = 2.257, p = .521$, root mean standard error of approximation (RMSEA) = .006, confirmatory fit index (CFI) = 1.00, standardized root mean square residual (SRMR) = .007. Standardized coefficients are presented. Risk profiles are modeled as dummy variables, with the low-risk profile serving as the reference category. Only significant paths at the $p < .05$ level are displayed. All implied indirect pathways of risk profile on outcomes are statistically significant. Also, all within-time covariances for investments-responsivity and all cognitive outcomes were positive and significant (see the Appendix in the online supplemental materials).

collectivist societies (Rothbaum, Weisz, Pott, Miyake, & Morelli, 2000), evidence suggests that the responsivity dimension, per se, is important across multiple, diverse groups. For example, Bornstein and colleagues describe a core set of interactional behaviors that appear to generalize across cultural contexts, including responsiveness to infant cues, sensitivity, and warmth (e.g., Bornstein, 2013; Bornstein & Cheah, 2006). Moreover, there is cross-cultural predictive validity of such behavior, whereby responsivity is associated with positive development in Western and non-Western groups (Gunning et al., 2004; Huang, Lewin, Mitchell, & Zhang, 2012; Kermani & Brenner, 2000).

When conceptualized as maternal responsivity during real-time, dyadic exchanges, the current study suggests that family processes do, indeed, provide an indirect link via contextual risk and school readiness among children in families characterized by differing patterns of adversity. Responsive exchanges appear to be *positive and development-enhancing social experiences*, associated with the emergence of skills that prepare children for school success. Findings from the current study replicate both Bernier and colleagues (2010) and Vernon-Feagans (2013), suggesting that responsivity is particularly important for the development of EF. The present study demonstrated that there was *not* a significant pathway between family investments and EF, though there *was* a significant effect of responsivity. Consistent with previous research, highly attuned exchanges appear to facilitate the emergence and internalization of self-regulation. Through such experiences, children become able to negotiate tasks requiring response inhibition and set shifting, in addition to the coordination of intentional and cooperative interactions that are linked to the emergence of social cognition and ToM (Tomasello & Carpenter, 2007). Concurrently, responsivity is also linked to skills in the areas of mathematics, reading, and language capacity when children are beginning school (Benson et al., 2013; Fitzpatrick, McKinnon, Blair, & Willoughby, 2014). Unfortunately, these experiences were less likely in the low-SES immigrant risk and low-SES multilevel risk homes, putting these children at a relative disadvantage in school readiness. Thus, the present study provides evidence that responsivity is, in fact, an important mechanism connecting risk and child cognition for a variety of outcomes at school entry. As others have argued, environmental exposures result in the *experiential canalization* of particular cognitive functions, and, although reversals are possible, the endurance of these individual differences are strong (Blair & Raver, 2012). It should be noted that the indirect effects via responsivity (like investments) are of small magnitude (Mistry et al., 2010). Thus, the outlined pathways are important, though they are likely not the only pathways linking adversity and cognition.

Similar to Lanza and colleagues (2010), the present study identified a particular group of families comprised of mothers who had developmental histories of substantial risk and adversity in their birth homes. However, unlike Lanza and colleagues, these mothers were *not* currently single and their children were *not* experiencing

Table 3
Summary of Mediation Effect Sizes as Partially Standardized Indirect Effects and Wald Tests Examining Equivalence of Effects through Investments Versus Responsivity

Risk profile 2 months	Outcome 4.5 years	Mediator 18 months	ab_{ps} (*SE*)	*p*	*Wald*	*p*
Low-SES immigrant risk	Achievement	Investments	−.09 (.04)	.02*	.16	.69
		Responsivity	−.07 (.03)	.05*		
	Vocabulary	Investments	−.12 (.04)	.01*	.84	.36
		Responsivity	−.07 (.03)	.04*		
	Theory of mind	Investments	−.10 (.04)	.02*	.18	.67
		Responsivity	−.07 (.04)	.04*		
	Executive function	Investments	−.04 (.03)	.21	.55	.46
		Responsivity	−.07 (.03)	.03*		
Low-SES multilevel risk	Achievement	Investments	−.16 (.06)	.01*	.08	.78
		Responsivity	−.14 (.06)	.02*		
	Vocabulary	Investments	−.21 (.06)	.00*	2.81	.09
		Responsivity	−.14 (.05)	.01*		
	Theory of mind	Investments	−.17 (.07)	.01*	.10	.75
		Responsivity	−.15 (.05)	.01*		
	Executive function	Investments	−.07 (.06)	.21	.99	.31
		Responsivity	−.15 (.05)	.01*		
Abuse history	Achievement	Investments	.03 (.03)	.18	.29	.59
		Responsivity	.02 (.02)	.33		
	Vocabulary	Investments	.04 (.03)	.13	.62	.43
		Responsivity	.02 (.03)	.56		
	Theory of mind	Investments	.04 (.03)	.13	.34	.56
		Responsivity	.02 (.03)	.56		
	Executive function	Investments	.02 (.02)	.33	.01	.95
		Responsivity	.02 (.03)	.60		

Note. Reference category is the low-risk group. Indirect effects are calculated using the Sobel method. Partially standardized indirect effects (ab_{ps}) are given as the product of the unstandardized indirect pathways divided by the standard deviation of the outcome variable (a^*b/σ_Y). Wald tests examine the null hypothesis that the investments and responsivity indirect effect sizes for a particular outcome and risk group are equivalent. Model covariates: female child, age, maternal language, birth weight. *SE* = standard error; SES = socioeconomic status.
$^* p < .05$.

difficulties. In fact, children in these homes had higher EF, and similar achievement, vocabulary, and ToM, compared with families in the low-risk profile. There was a direct effect linking membership in the abuse history profile and higher EF, over and above the effects of the family investment and process pathways. Intergenerational studies have demonstrated that individuals who grow up in dysfunctional homes but marry into healthy relationships break cycles of deprivation (Conger, Schofield, & Neppl, 2012). Indeed, functional paternal and sibling relationships can promote positive development, over and above the effects of positive interactions with maternal caretakers (J. Dunn, Davies, O'Connor, & Sturgess, 2000; Gass, Jenkins, & Dunn, 2007). Moreover, it is possible that genetic factors have created phenotypic similarity among high-functioning children and mothers who are able to overcome adversity (Doyle et al., 2005).

Table 4
Direct Effects of Risk Profile on Cognitive Outcomes from Path Model

Direct Effect	β *SE*	*p*
Low-SES Immigrant → Achievement	.01 (.06)	.994
Low-SES Immigrant → Vocabulary	−.07 (.06)	.254
Low-SES Immigrant → Theory of mind	.03 (.06)	.638
Low-SES Immigrant → Executive function	−.02 (.06)	.771
Low-SES Multilevel → Achievement	−.02 (.06)	.767
Low-SES Multilevel → Vocabulary	−.02 (.06)	.763
Low-SES Multilevel → Theory of mind	.06 (.05)	.277
Low-SES Multilevel → Executive function	.02 (.05)	.680
Abuse History → Achievement	.01 (.04)	.924
Abuse History → Vocabulary	.06 (.04)	.172
Abuse History → Theory of mind	.05 (.05)	.281
Abuse History → Executive function	.10 (.05)	.034

Note. *SE* = standard error; SES = socioeconomic status.

This prospective birth-cohort study employed longitudinal, multimethod, and multi-informant data, including mother report, interviewer and videotaped observation, assessment, and national census data. Findings operated across assessment modalities, reducing the likelihood of shared method variance bias. That being said, there are a few limitations that future research should address. First, a unique group of immigrants was identified to be at risk. However, the current study did not address other cultural processes, such as beliefs, values, customs, and parental cognitions. Although links between parenting and child outcome did not vary significantly between multilevel-risk and low-SES immigrant families, it is possible that different mediation processes would have been uncovered if other cultural measures were included. Also, the meaning of certain constructs may differ across cultural groups (Delpit, 2006). Second, the sample was more educated than the population and only assessed in English, reducing generalizability. Given that the sampling resulted in diminished variation on risk, yet substantial variance in child outcomes was explained (between 18% and 26%), it is possible that greater representativeness may have yielded even stronger effects. That being said, the inclusion of non-English-speaking families could have yielded an additional latent class that may or may not have been linked to family mediators. Third, family mediation was considered for two variables: (a) maternal responsivity, and (b) family investment. Both of these constructs were conceptualized as overarching mechanisms, though there may have been differential mediation via the subcomponents of these mechanisms (e.g., investments being comprised of both order vs. chaos *and* learning materials). Future research might consider differential mediation effects via these subcomponents. Fourth, all child outcomes were measured contemporaneously, preventing an examination of the temporal primacy of certain skills. Fifth, given the large number of variables used in the LCA, it is important for different samples to replicate the groups identified in the present study. Finally, we did not include any measurement surrounding the family process contributions of fathers or other caregivers (e.g., grandparents, extended family, nonbiological caregivers). The field would benefit from future studies that address these gaps.

From a practice and policy perspective, our results add to nonexperimental and experimental designs that articulate the centrality of reducing poverty for the promotion of early developmental success. In accordance with the outlined pathways throughout the early years, initiatives to promote school success are most effective when they occur at the earliest stages of life (Heckman, 2006). Furthermore, our results identify a subgroup of immigrant families whose children are struggling and are in particular need for resettlement assistance, particularly in the economic realm. These families may be slipping through the cracks of social safety nets because of the absence of broad-band risk, despite the presence of economic and neighborhood stress. In other words, children within these families may be falling behind, partially because of undetected and economically influenced factors that are happening inside the home. It is possible that these families are *not* coming to the attention of educators and social service workers, given that there are other families who may be more visibly at risk. Moving forward, the changing immigration landscape necessitates the development of innovative ways for providing access to culturally sensitive screening, assessment, and interventions, in addition to economic programming. Efforts should include programs that promote school readiness via the enhancement of maternal responsivity (e.g., Landry et al., 2006), even if there does not appear to be systematic family difficulties or a long history of dysfunction.

References

Abidin, R. R. (1990). Introduction to the special issue: The stresses of parenting. *Journal of Clinical Child Psychology, 19,* 298–301. http://dx.doi.org/10.1207/s15374424jccp1904_1

Alba, R., & Foner, N. (2015). *Strangers no more: Immigration and the challenges of integration in North America and Western Europe*. Princeton, NJ: Princeton University Press. http://dx.doi.org/10.1515/9781400865901

Almedom, A. M. (2005). Social capital and mental health: An interdisciplinary review of primary evidence. *Social Science & Medicine, 61,* 943–964. http://dx.doi.org/10.1016/j.socscimed.2004.12.025

Asparouhov, T., & Muthén, B. (2012). Using Mplus TECH11 and TECH14 to test the number of latent classes. *Mplus Web Notes No. 14*. Retrieved from https://www.statmodel.com/examples/webnotes/webnote14.pdf

Astington, J. W., & Jenkins, J. M. (1999). A longitudinal study of the relation between language and theory-of-mind development. *Developmental Psychology, 35,* 1311–1320. http://dx.doi.org/10.1037/0012-1649.35.5.1311

Benson, J. E., Sabbagh, M. A., Carlson, S. M., & Zelazo, P. D. (2013). Individual differences in executive functioning predict preschoolers' improvement from theory-of-mind training. *Developmental Psychology, 49,* 1615–1627. http://dx.doi.org/10.1037/a0031056

Bernier, A., Carlson, S. M., Deschênes, M., & Matte-Gagné, C. (2012). Social factors in the development of early executive functioning: A closer look at the caregiving environment. *Developmental Science, 15,* 12–24. http://dx.doi.org/10.1111/j.1467-7687.2011.01093.x

Bernier, A., Carlson, S. M., & Whipple, N. (2010). From external regulation to self-regulation: Early parenting precursors of young children's executive functioning. *Child Development, 81,* 326–339. http://dx.doi.org/10.1111/j.1467-8624.2009.01397.x

Blair, C., & Raver, C. C. (2012). Child development in the context of adversity: Experiential canalization of brain and behavior. *American Psychologist, 67,* 309–318. http://dx.doi.org/10.1037/a0027493

Boivin, M., & Bierman, K. L. (2014). *Promoting school readiness and early learning: Implications of developmental research for practice*. New York, NY: Guilford Press.

Bornstein, M. H. (2013). Cross-cultural perspectives on parenting. *International Perspectives on Psychological Science, 2,* 359–369.

Bornstein, M. H., & Cheah, C. S. (2006). The place of "culture and parenting" in the ecological contextual perspective on developmental science. In K. H. Rubin & O. B. Chung (Eds.), *Parenting beliefs, behaviors, and parent-child relations: A cross-cultural perspective* (pp. 3–33). New York, NY: Taylor & Francis.

Boyle, M. H., Offord, D. R., Racine, Y., Fleming, J. E., Szatmari, P., & Sanford, M. (1993). Evaluation of the revised Ontario child health study scales. *Child Psychology & Psychiatry & Allied Disciplines, 34,* 189–213. http://dx.doi.org/10.1111/j.1469-7610.1993.tb00979.x

Bradley, R. H. (1994). The Home Inventory: Review and reflections. *Advances in Child Development and Behavior, 25,* 241–288. http://dx.doi.org/10.1016/S0065-2407(08)60054-3

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology, 53,* 371–399. http://dx.doi.org/10.1146/annurev.psych.53.100901.135233

Brody, G. H., Yu, T., Chen, Y. F., Kogan, S. M., Evans, G. W., Beach, S. R., . . . Philibert, R. A. (2013). Cumulative socioeconomic status risk, allostatic load, and adjustment: A prospective latent profile analysis with contextual and genetic protective factors. *Developmental Psychology, 49,* 913–927. http://dx.doi.org/10.1037/a0028847

Carlson, S. M., Moses, L. J., & Claxton, L. J. (2004). Individual differences in executive functioning and theory of mind: An investigation of inhibitory control and planning ability. *Journal of Experimental Child Psychology, 87,* 299–319. http://dx.doi.org/10.1016/j.jecp.2004.01.002

CIC News. (2015). *U.S. and Canadian immigration policies: A comparison.* Retrieved from http://www.cicnews.com/2015/08/canadian-immigration-policies-comparison-085861.html

Claussen, A. H., & Crittenden, P. M. (2000). *Maternal sensitivity*. New York, NY: Cambridge University Press.

Coll, C. G., & Marks, A. K. (Eds.). (2012). *The immigrant paradox in children and adolescents: Is becoming American a developmental risk?* Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/13094-000

Conger, R. D., Conger, K. J., & Martin, M. J. (2010). Socioeconomic Status, Family Processes, and Individual Development. *Journal of Marriage and Family, 72,* 685–704. http://dx.doi.org/10.1111/j.1741-3737.2010.00725.x

Conger, R. D., Schofield, T. J., & Neppl, T. K. (2012). Intergenerational continuity and discontinuity in harsh parenting. *Parenting: Science and Practice, 12,* 222–231. http://dx.doi.org/10.1080/15295192.2012.683360

Copeland, W., Shanahan, L., Costello, E. J., & Angold, A. (2009). Configurations of common childhood psychosocial risk factors. *Journal of Child Psychology and Psychiatry, 50,* 451–459. http://dx.doi.org/10.1111/j.1469-7610.2008.02005.x

Creese, G., & Wiebe, B. (2012). 'Survival employment': Gender and deskilling among African immigrants in Canada. *International Migration, 50,* 56–76. http://dx.doi.org/10.1111/j.1468-2435.2009.00531.x

Deater-Deckard. K., Pike, A., Petrill, S. A., Cutting, A. L., Hughes, C., & O'Connor, T. G. (2001). Nonshared environmental processes in social-emotional development: An observational study of identical twin differences in the preschool period. *Developmental Science, 4,* F1–F6. http://dx.doi.org/10.1111/1467-7687.00157

Deater-Deckard. K., Pylas. M. V., & Petrill, S. A. (1997). *Parent-Child Interaction System (PARCHISY).* London, UK: Institute of Psychiatry.

Delpit, L. D. (2006). *Other people's children: Cultural conflict in the classroom.* New York, NY: The New Press.

de Rosnay, M.. & Murray, L. (2012). Maternal care as the central environmental variable. In L. C. Mayes & M. Lewis (Ed.), *The Cambridge handbook of environment in human development* (pp. 58–82). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139016827.005

Doyle. A. E., Faraone, S. V., Seidman. L. J., Willcutt. E. G., Nigg, J. T., Waldman, I. D., . . . Biederman. J. (2005). Are endophenotypes based on measures of executive functions useful for molecular genetic studies of ADHD? *Journal of Child Psychology and Psychiatry, 46,* 774–803. http://dx.doi.org/10.1111/j.1469-7610.2005.01476.x

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43,* 1428–1446. http://dx.doi.org/10.1037/0012-1649.43.6.1428

Duncan, G. J., & Magnuson, K. (2012). Socioeconomic status and cognitive functioning: Moving from correlation to causation. *WIREs Cognitive Science, 3,* 377–386. http://dx.doi.org/10.1002/wcs.1176

Dunn, J., Davies, L. C., O'Connor, T. G., & Sturgess, W. (2000). Parents' and partners' life course and family experiences: Links with parent-child relationships in different family settings. *Journal of Child Psychology and Psychiatry, 41,* 955–968. http://dx.doi.org/10.1111/1469-7610.00684

Dunn, L. M., & Dunn, D. M. (2007). *PPVT-4: Peabody Picture Vocabulary Test.* Minneapolis, MN: Pearson Assessments.

Ensor, R., Spencer, D., & Hughes, C. (2011). "You feel sad?" Emotion understanding mediates effects of verbal ability and mother–child mutuality on prosocial behaviors: Findings from 2 years to 4 years. *Social Development, 20,* 93–110. http://dx.doi.org/10.1111/j.1467-9507.2009.00572.x

Evans, G. W., & Kim, P. (2012). Childhood poverty and young adults' allostatic load: The mediating role of childhood cumulative risk exposure. *Psychological Science, 23,* 979–983. http://dx.doi.org/10.1177/0956797612441218

Fernyhough, C. (2008). Getting Vygotskian about theory of mind: Mediation, dialogue, and the development of social understanding. *Developmental Review, 28,* 225–262. http://dx.doi.org/10.1016/j.dr.2007.03.001

Fitzpatrick, C., McKinnon, R. D., Blair, C. B., & Willoughby, M. T. (2014). Do preschool executive function skills explain the school readiness gap between advantaged and disadvantaged children? *Learning and Instruction, 30,* 25–31. http://dx.doi.org/10.1016/j.learninstruc.2013.11.003

Gass, K., Jenkins, J., & Dunn, J. (2007). Are sibling relationships protective? A longitudinal study. *Journal of Child Psychology and Psychiatry, 48,* 167–175. http://dx.doi.org/10.1111/j.1469-7610.2006.01699.x

Gazso, A., & Waldron, I. (2009). Fleshing out the racial undertones of poverty for Canadian women and their families: Re-envisioning a critical integrative approach. *Atlantis: Critical Studies in Gender, Culture &. Social Justice, 34,* 132–141.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576. http://dx.doi.org/10.1146/annurev.psych.58.110405.085530

Gunning, M., Conroy, S., Valoriani, V., Figueiredo, B., Kammerer, M. H., Muzik, M., . . . TCS-PND Group. (2004). Measurement of mother-infant interactions and the home environment in a European setting: Preliminary results from a cross-cultural study. *The British Journal of Psychiatry, 184,* s38–s44. http://dx.doi.org/10.1192/bjp.184.46.s38

Guo, G., & Harris, K. M. (2000). The mechanisms mediating the effects of poverty on children's intellectual development. *Demography, 37,* 431–447. http://dx.doi.org/10.1353/dem.2000.0005

Hart, B., & Risley, T. R. (1992). American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. *Developmental Psychology, 28,* 1096–1105. http://dx.doi.org/10.1037/0012-1649.28.6.1096

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science, 312,* 1900–1902. http://dx.doi.org/10.1126/science.1128898

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development, 74,* 1368–1378. http://dx.doi.org/10.1111/1467-8624.00612

Huang, Z. J., Lewin, A., Mitchell, S. J., & Zhang, J. (2012). Variations in the relationship between maternal depression, maternal sensitivity, and child attachment by race/ethnicity and nativity: Findings from a nationally representative cohort study. *Maternal and Child Health Journal, 16,* 40–50. http://dx.doi.org/10.1007/s10995-010-0716-2

Hughes, C. (1998). Executive function in preschoolers: Links with theory of mind and verbal ability. *British Journal of Developmental Psychology, 16,* 233–253. http://dx.doi.org/10.1111/j.2044-835X.1998.tb00921.x

Hughes, C., & Ensor, R. (2005). Executive function and theory of mind in 2 year olds: A family affair? *Developmental Neuropsychology, 28,* 645–668. http://dx.doi.org/10.1207/s15326942dn2802_5

Jansen, P. W., Raat, H., Mackenbach, J. P., Jaddoe, V. W., Hofman, A., van Oort, F. V., . . . Tiemeier, H. (2010). National origin and behavioural problems of toddlers: The role of family risk factors and maternal immigration characteristics. *Journal of Abnormal Child Psychology, 38,* 1151–1164. http://dx.doi.org/10.1007/s10802-010-9424-z

Janus, M., & Duku, E. (2007). The school entry gap: Socioeconomic, family, and health factors associated with children's school readiness to learn. *Early Education and Development, 18,* 375–403. http://dx.doi.org/10.1080/10409280701610796a

Kerig, P. K. (1996). Assessing the links between interparental conflict and child adjustment: The conflicts and problem-solving scales. *Journal of Family Psychology, 10,* 454–473. http://dx.doi.org/10.1037/0893-3200.10.4.454

Kermani, H., & Brenner, M. E. (2000). Maternal scaffolding in the child's zone of proximal development across tasks: Cross-cultural perspectives. *Journal of Research in Childhood Education, 15,* 30–52. http://dx.doi.org/10.1080/02568540009594774

Koury, A. S., & Votruba-Drzal, E. (2014). School readiness of children from immigrant families: Contributions of region of origin, home, and childcare. *Journal of Educational Psychology, 106,* 268–288. http://dx.doi.org/10.1037/a0034374

Kowlessar, N. M., Jiang, H. J., & Steiner, C. (2013). *Hospital stays for newborns, 2011. HCUP Statistical Brief #163.* Rockville, MD: Agency for Healthcare Research and Quality.

Landry, S. H., Smith, K. E., & Swank, P. R. (2006). Responsive parenting: Establishing early foundations for social, communication, and independent problem-solving skills. *Developmental Psychology, 42,* 627–642. http://dx.doi.org/10.1037/0012-1649.42.4.627

Lanza, S. T., & Rhoades, B. L. (2013). Latent class analysis: An alternative perspective on subgroup analysis in prevention and treatment. *Prevention Science, 14,* 157–168. http://dx.doi.org/10.1007/s11121-011-0201-1

Lanza, S. T., Rhoades, B. L., Nix, R. L., & Greenberg, M. T., & Conduct Problems Prevention Research Group. (2010). Modeling the interplay of multilevel risk factors for future academic and behavior problems: A person-centered approach. *Development and Psychopathology, 22,* 313–335. http://dx.doi.org/10.1017/S0954579410000088

Lee, D., Brooks-Gunn, J., McLanahan, S. S., Notterman, D., & Garfinkel, I. (2013). The Great Recession, genetic sensitivity, and maternal harsh parenting. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 110,* 13780–13784. http://dx.doi.org/10.1073/pnas.1312398110

Linver, M. R., Brooks-Gunn, J., & Kohen, D. E. (2002). Family processes as pathways from income to young children's development. *Developmental Psychology, 38,* 719–734. http://dx.doi.org/10.1037/0012-1649.38.5.719

Matias, C., Scott, S., & O'Connor, T. G. (2006). *Coding of Attachment-Related Parenting (CARP).* London, UK: Institute of Psychiatry, King's College.

McGuire, J. B. (1997). The reliability and validity of a questionnaire describing neighborhood characteristics relevant to families and young children living in urban areas. *Journal of Community Psychology, 25,* 551–566. http://dx.doi.org/10.1002/(SICI)1520-6629(199711)25:6<551::AID-JCOP5>3.0.CO:2-S

Meunier, J. C., Boyle, M., O'Connor, T. G., & Jenkins, J. M. (2013). Multilevel mediation: Cumulative contextual risk, maternal differential treatment, and children's behavior within families. *Child Development, 84,* 1594–1615. http://dx.doi.org/10.1111/cdev.12066

Mistry, R. S., Benner, A. D., Biesanz, J. C., Clark, S. L., & Howes, C. (2010). Family and social risk, and parental investments during the early childhood years as predictors of low-income children's school readiness outcomes. *Early Childhood Research Quarterly, 25,* 432–449. http://dx.doi.org/10.1016/j.ecresq.2010.01.002

Mistry, R. S., Biesanz, J. C., Chien, N., Howes, C., & Benner, A. D. (2008). Socioeconomic status, parental investments, and the cognitive and behavioral outcomes of low-income children from immigrant and native households. *Early Childhood Research Quarterly, 23,* 193–212. http://dx.doi.org/10.1016/j.ecresq.2008.01.002

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Author.

National Institute of Child Health and Human Development Early Child Care Research Network. (2005). Duration and developmental timing of poverty and children's cognitive and social development from birth through third grade. *Child Development, 76,* 795–810. http://dx.doi.org/10.1111/j.1467-8624.2005.00878.x

Noble, K. G., Norman, M. F., & Farah, M. J. (2005). Neurocognitive correlates of socioeconomic status in kindergarten children. *Developmental Science, 8,* 74–87. http://dx.doi.org/10.1111/j.1467-7687.2005.00394.x

Perreira, K. M., Chapman, M. V., & Stein, G. L. (2006). Becoming an American parent: Overcoming challenges and finding strength in a new immigrant Latino community. *Journal of Family Issues, 27,* 1383–1414. http://dx.doi.org/10.1177/0192513X06290041

Picot, G. (2004). The deteriorating economic welfare of Canadian immigrants. *Canadian Journal of Urban Research, 13,* 25–45.

Pinderhughes, E. E., Dodge, K. A., Bates, J. E., Pettit, G. S., & Zelli, A. (2000). Discipline responses: Influences of parents' socioeconomic status, ethnicity, beliefs about parenting, stress, and cognitive-emotional processes. *Journal of Family Psychology, 14,* 380–400. http://dx.doi.org/10.1037/0893-3200.14.3.380

Pratt, M. E., McClelland, M. M., Swanson, J., & Lipscomb, S. T. (2016). Family risk profiles and school readiness: A person-centered approach. *Early Childhood Research Quarterly, 36,* 462–474. http://dx.doi.org/10.1016/j.ecresq.2016.01.017

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods, 16,* 93–115. http://dx.doi.org/10.1037/a0022658

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385–401. http://dx.doi.org/10.1177/014662167700100306

Raviv, T., Kessenich, M., & Morrison, F. J. (2004). A mediational model of the association between socioeconomic status and three-year-old language abilities: The role of parenting factors. *Early Childhood Research Quarterly, 19,* 528–547. http://dx.doi.org/10.1016/j.ecresq.2004.10.007

Reed, M. A., Pien, D. L., & Rothbart, M. K. (1984). Inhibitory self-control in preschool children. *Merrill-Palmer Quarterly, 30,* 131–147.

Rothbaum, F., Weisz, J., Pott, M., Miyake, K., & Morelli, G. (2000). Attachment and culture. Security in the United States and Japan. *American Psychologist, 55,* 1093–1104. http://dx.doi.org/10.1037/0003-066X.55.10.1093

Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science, 277,* 918–924. http://dx.doi.org/10.1126/science.277.5328.918

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75,* 417–453. http://dx.doi.org/10.3102/00346543075003417

Sobel, M. E. (1987). Direct and indirect effects in linear structural equation models. *Sociological Methods & Research, 16,* 155–176. http://dx.doi.org/10.1177/0049124187016001006

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360–407. http://dx.doi.org/10.1598/RRQ.21.4.1

Statistics Canada. (2006). *2006 Census of Population.* Retrieved from http://www12.statcan.gc.ca/census-recensement/2006/index-eng.cfm

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9,* 1–19.

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science, 10,* 121–125. http://dx.doi.org/10.1111/j.1467-7687.2007.00573.x

UNHCR. (2015). *Global trends: Forced displacement in 2015.* Retrieved from http://www.unhcr.org/en-us/statistics/unhcrstats/576408cd7/unhcr-global-trends-2015.html

Vernon-Feagans, L. (2013). Cumulative risk and its relation to parenting and child outcomes at 36 months. *Monographs of the Society for Research in Child Development, 78,* 66–91. http://dx.doi.org/10.1111/mono.12051

Vygotsky, L. (1978). *Mind in society.* Cambridge, MA: Harvard University Press.

Walks, R., & Bourne, L. S. (2006). Ghettos in Canada's cities? Racial segregation, ethnic enclaves and poverty concentration in Canadian urban areas. *The Canadian Geographer, 50,* 273–297. http://dx.doi.org/10.1111/j.1541-0064.2006.00142.x

Walsh, C. A., MacMillan, H. L., Trocmé, N., Jamieson, E., & Boyle, M. H. (2008). Measurement of victimization in adolescence: Development and validation of the Childhood Experiences of Violence Questionnaire. *Child Abuse & Neglect, 32,* 1037–1057. http://dx.doi.org/10.1016/j.chiabu.2008.05.003

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75,* 523–541. http://dx.doi.org/10.1111/j.1467-8624.2004.00691.x

Wilkinson, R., & Pickett, K. (2009). *The spirit level: Why equality is better for everyone.* London, UK: Allen Lane.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson, Tests of Achievement, Third Edition.* Itasca, IL: Riverside.

Yeung, W. J., Linver, M. R., & Brooks-Gunn, J. (2002). How money matters for young children's development: Parental investment and family processes. *Child Development, 73,* 1861–1879. http://dx.doi.org/10.1111/1467-8624.t01-1-00511

Zambrana, I. M., Ystrom, E., & Pons, F. (2012). Impact of gender, maternal education, and birth order on the development of language comprehension: A longitudinal study from 18 to 36 months of age. *Journal of Developmental and Behavioral Pediatrics, 33,* 146–155.

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature Protocols, 1,* 297–301. http://dx.doi.org/10.1038/nprot.2006.46

Zimmerman, C., Kiss, L., & Hossain, M. (2011). Migration and health: A framework for 21st century policy-making. *PLoS Medicine, 8*(5), e1001034. http://dx.doi.org/10.1371/journal.pmed.1001034

Received December 8, 2015
Revision received January 24, 2017
Accepted February 21, 2017 ■

Call for Papers
A Focused Collection of Qualitative Studies in the Psychological Sciences: Reasoning and Participation in Formal and Informal Learning Environments

Journal of Educational Psychology
Guest Editors: Tanner LeBaron Wallace and Eric Kuo

Reasoning and participation are two central topics of education research in the psychological sciences. Understanding the mechanisms that govern thought and reasoning has long been a core enterprise of educational psychology and, over time, more modern views on learning have promoted participation as a key feature for research—either as a facilitator of learning, a practice to be learned, or as an operationalization of learning itself.

We are pleased to announce a focused collection highlighting qualitative studies of reasoning and participation in formal and informal learning environments. By inviting studies incorporating qualitative methods, we aim to complement the experimental and longitudinal statistical research on these topics that is typically published in this journal. We encourage submission of papers focused on the following (or closely related) topics:

- Student reasoning and/or participation in novel learning environments or activities
- The relations between student reasoning, motivation, identity, and participation
- Student perceptions and meaning-making during participatory experiences
- Dynamic models of student reasoning that are grounded in data
- Explanatory accounts for how and why participation is successful (or not)
- Identifying new goals or targeted outcomes for reasoning or participation

We especially welcome qualitative studies that demonstrate the possibilities for unique discovery afforded by inductive analysis of rich data sources (e.g., real-time recordings of student reasoning, participation, discourse, and physical action, students' meaning-making anchored to particular interactions experienced). This collection will highlight the benefits of qualitative methods for extending and deepening theoretical and empirical understandings of reasoning and participation in both formal and informal learning environments.

The deadline for manuscript submissions is **March 1, 2018**. We invite authors to contact the Guest Editors of this collection, Tanner LeBaron Wallace (twallace@pitt.edu) and Eric Kuo (erickuo@pitt.edu), for discussion on how to maximize alignment between their submissions and this focused collection, though it is not required. Please follow both APA guidelines as well as specific submission criteria for the journal. When submitting manuscripts, please also indicate your intent to submit to this focused collection in the required cover letter.

All manuscripts must be submitted electronically at http://www.editorialmanager.com/edu. In the submission portal, please select the article type "Special Section: Reasoning & Participation – Qualitative." For more information on the *Journal of Educational Psychology*, please visit http://www.apa.org/pubs/journals/edu/.

Instructions to Authors
Journal of Educational Psychology

www.apa.org/pubs/journals/edu

The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Manuscript preparation. Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (6th ed.). Manuscripts may be copyedited for bias-free language (see pp. 70–77 of the *Publication Manual*). Formatting instructions (all copy must be double-spaced) and instructions on the preparation of tables, figures, references, metrics, and abstracts appear in the *Manual*. For APA's Checklist for Manuscript Submission, see www.apa.org/pubs/journals/edu. **Abstract and keywords.** All manuscripts must include an abstract containing a maximum of 250 words typed on a separate page. After the abstract, please supply up to five keywords or brief phrases. **References.** References should be listed in alphabetical order. Each listed reference should be cited in text, and each text citation should be listed in the References. Basic formats are as follows:

Hughes, G., Desantis, A., & Waszak, F. (2013). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychologial Bulletin, 139,* 133–151. http://dx.doi.org/10.1037/a0028566

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.

Gill, M. J., & Sypher, B. D. (2009). Workplace incivility and organizational trust. In P. Lutgen-Sandvik & B. D. Sypher (Eds.), *Destructive organizational communication: Processes, consequences, and constructive ways of organizing* (pp. 53–73). New York, NY: Taylor & Francis.

Adequate description of participants is critical to the science and practice of educational psychology; this allows readers to assess the results, determine generalizability of findings, and make comparisons in replications, extensions, literature reviews, or secondary data analyses. Authors should see guidelines for sample–subject description in the *Manual*. Appropriate indexes of effect size or strength of relationship should be incorporated in the results section of the manuscript (see p. 34 of the *Manual*). Information that allows the reader to assess not only the significance but also the magnitude of the observed effects or relationships clarifies the importance of the findings. **Figures.** Graphics files are welcome if supplied in TIFF or EPS format. APA's policy on publication of color figures is available at http://www.apa.org/pubs/authors/instructions.aspx?item=6.

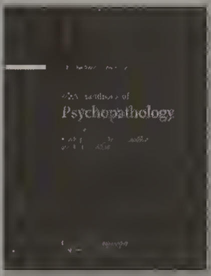
Publication policies. APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. APA policy regarding posting articles on the Internet may be found at www.apa.org/pubs/authors/posting.aspx. In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 8.13). As this is a primary journal that publishes original material only, APA policy prohibits publication of any manuscript or data that have already been published in whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data on which their article depends. In addition, APA Ethical Principles specify that "after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 8.14). Authors must have available their data throughout the editorial review process and for at least 5 years after the date of publication.

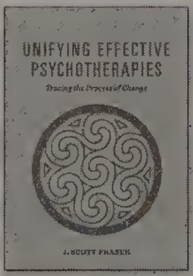
Masked review policy. The *Journal* has a masked review policy, which means that the identities of both authors and reviewers are masked. Every effort should be made by the authors to see that the manuscript itself contains no clues to their identities. Authors should never use first person (*I, my, we, our*) when referring to a study conducted by the author(s) or when doing so reveals the authors' identities, e.g., "in our previous work, Johnson et al., 1998 reported that . . .." Instead, references to the authors' work should be in third person, e.g., "Johnson et al. (1998) reported that . . .." The authors' institutional affiliations should also be masked in the manuscript. Authors submitting manuscripts are required to include in the cover letter the title of the manuscript along with all authors' names and institutional affiliations. However, the first page of the manuscript should omit the authors' names and affiliations, but should include the title of the manuscript and the date it is submitted. Responsibility for masking the manuscript rests with the authors; manuscripts will be returned to the author if not appropriately masked. If the manuscript is accepted, authors will be asked to make changes in wording so that the paper is no longer masked. Authors are required to state in writing that they have complied with APA ethical standards in the treatment of their sample, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained at www.apa.org/ethics/ or by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242. APA requires authors to reveal any possible conflict of interest in the conduct and reporting of research (e.g., financial interests in a test procedure, funding by pharmaceutical companies for drug research). Authors of accepted manuscripts will be required to transfer copyright to APA.

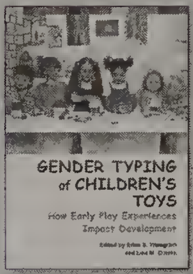
Permissions. Authors of accepted papers must obtain and provide to the editor on final acceptance all necessary permissions to reproduce in print and electronic form any copyrighted work, including test materials (or portions thereof), photographs, and other graphic images (including those used as stimuli in experiments). On advice of counsel, APA may decline to publish any image whose copyright status is unknown.

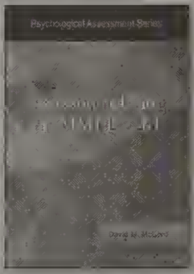
Supplemental materials. APA can place supplementary materials online, which will be available via the published article in the PsycARTICLES database. To submit such materials, please see www.apa.org/pubs/authors/supp-material.aspx for details. Authors of accepted papers will be asked to work with the editor and production staff to provide supplementary materials as appropriate.

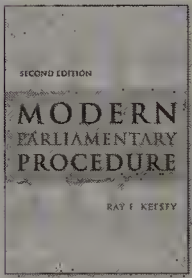
Submission. Authors should submit their manuscripts electronically via the Manuscript Submission Portal at www.apa.org/pubs/journals/edu/index.aspx (follow the link for submission under Instructions to Authors). General correspondence may be addressed to the editorial office at CJohnson@apa.org.

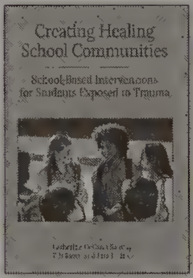

Psychopathology

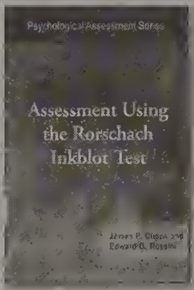

UNIFYING EFFECTIVE PSYCHOTHERAPIES
Tracing the Process of Change
J. SCOTT FRASER

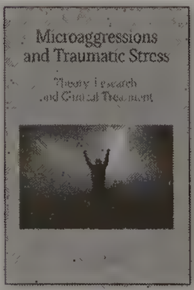

GENDER TYPING of CHILDREN'S TOYS
How Early Play Experiences Impact Development

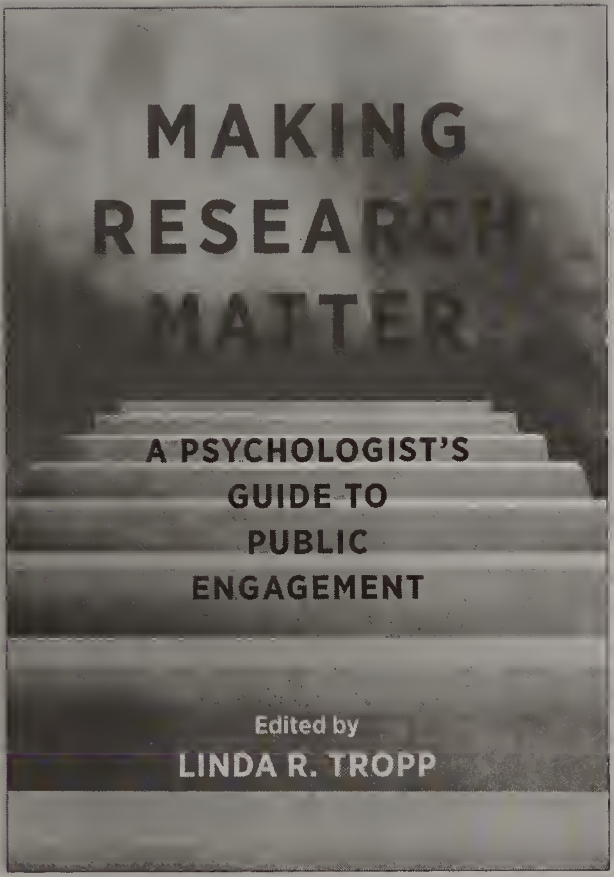

Psychological Assessment Series

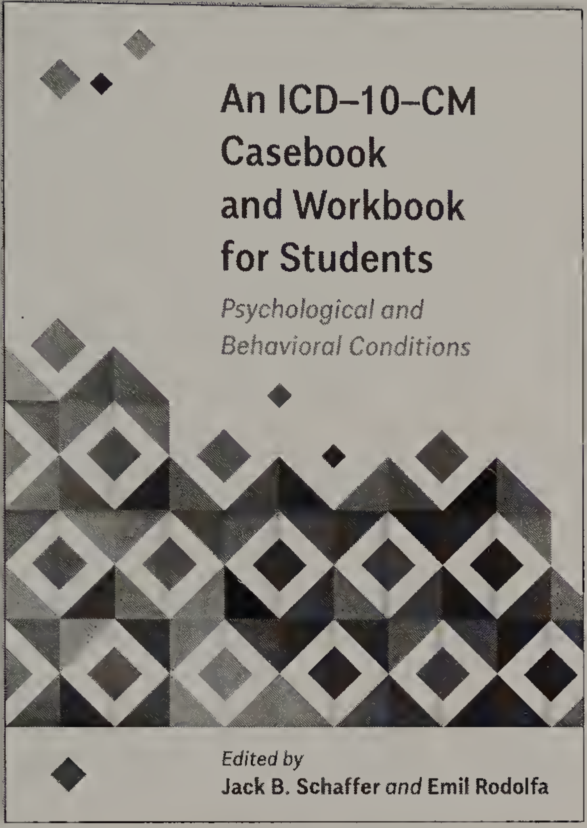

SECOND EDITION
MODERN PARLIAMENTARY PROCEDURE
RAY E. KEESEY


Creating Healing School Communities
School-Based Interventions for Students Exposed to Trauma


Psychological Assessment Series
Assessment Using the Rorschach Inkblot Test


Microaggressions and Traumatic Stress
Theory, Research and Clinical Treatment

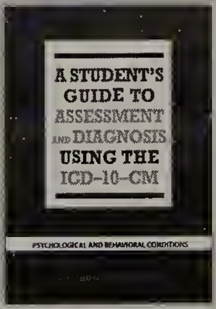

MAKING RESEARCH MATTER
A PSYCHOLOGIST'S GUIDE TO PUBLIC ENGAGEMENT
Edited by
LINDA R. TROPP



An ICD–10–CM Casebook and Workbook for Students

Psychological and Behavioral Conditions

Edited by
Jack B. Schaffer *and* Emil Rodolfa



A STUDENT'S GUIDE TO ASSESSMENT AND DIAGNOSIS USING THE ICD-10-CM

PSYCHOLOGICAL AND BEHAVIORAL CONDITIONS